Medical Concept Normalization in User-Generated Text

Master's Thesis

Emmanouil Manousogiannis

Medical Concept Normalization in User-Generated Text

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK DATA SCIENCE AND TECHNOLOGY

by

Emmanouil Manousogiannis
born in Heraklion, Greece

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
`http://wis.ewi.tudelft.nl`

Medical Concept Normalization in User-Generated Text

Author: Emmanouil Manousogiannis
Student id: 4727517
Email: E.Manousogiannis@student.tudelft.nl

Abstract

Online social networks have revolutionized the way people interact with each other nowadays. Users often share their experiences in various health - related topics like disease symptoms, drug treatments and other medical related issues in order to discuss with other patients dealing with similar conditions.

During the production of a new drug, important drug properties like possible Adverse Drug Reactions (ADRs) are monitored through a phase of clinical trials. However, due to various factors that can not be easily measured in those trials, patients can potentially experience adverse events that were not related to their treatment before, or were related to it in a much smaller frequency. Therefore, the automatic detection of Adverse Events in online networks is gaining an increasing popularity among researchers in the biomedical community, as it can offer a valuable complementary source of information, next to the traditional approaches of reporting those events to the corresponding Food and Drug Association. From an NLP perspective, this task poses a significant challenge as there is a large gap between the informal language used in social media and the formal medical language used to officially describe a medical concept. Moreover, there is an absence of large annotated datasets, as the manual labeling of an adverse effect mentions is a time-consuming and often ambiguous procedure which also requires some sort of medical expertise.

In this work we propose a novel machine learning approach to normalize Adverse Drug Effect mentions in user-generated text to a standard vocabulary from a medical Ontology. The evaluation results of the proposed model demonstrates a competitive performance among the current state of the art techniques, posing the potential feasibility of our model in the medical concept normalization domain.

Thesis Committee:

Chair:	Prof. dr. ir. Geert-Jan Houben, Faculty EEMCS, TUDelft
University supervisor:	Prof. dr. ir. Alessandro Bozzon, Industrial Design Engineering, TUDelft
Company supervisor:	Dr. Robert-Jan Sips, myTomorrows
Committee Member:	Dr. Ir. Jan Rellermeyer, Faculty EEMCS, TUDelft

Preface

This document is the result of my nine month Thesis project, as part of the Computer Science Master's Program at Delft University of Technology. I would like to take this opportunity to thank all members of the Thesis Committee and especially my supervisor Professor Alessandro Bozzon for his guidance throughout the whole project. In addition I would like to thank Robert-Jan Sips for giving me the opportunity to be part of a very innovative working environment and helping me take the right direction in my research with his valuable suggestions and feedback. Special thanks to Sepideh Mesbah and Selene Baez Santamaria for their extensive feedback and help on a daily basis as well as all my family members for their continuous support and help during my studies.

Emmanouil Manousogiannis
Delft, the Netherlands
September 4, 2019

Contents

List of Figures

Chapter 1

Introduction

Social media networks have become an almost inexhaustible source of people's opinions on a wide variety of different topics. As a result, there has already been a lot of research done in the field of extracting useful information from users' posts in social media or public forums for different domains. In the biomedical domain, extracting information that would allow us to monitor adverse events or indications of Drugs that are available in the market would be one extremely useful task for various reasons.

During the production of a new drug, the standard procedure before making it available to public is a preliminary phase of clinical trials , where a limited number of patients use the examined medicine in accordance with the instructions of a team of doctors. During this phase patients report the extent to which this drug improves their health, as well as any adverse reactions that this may cause. However, since the amount of people that take part in these trials is limited, it would be very useful to monitor patients experience on those drugs in a larger scale after the medicine is made available to everyone in the market. In most cases, this is not possible. Patients may report potential adverse drug reactions to the regulatory authorities such as the US Food and Drug Administration or the European Medicine Agency, however this kind of information is in most cases not reported unless the situation is critical. On the other hand, the rise of social media networks has led people discuss their health related issues in public and share their experiences. Twitter, Reddit and other famous public forums can be a large source of information for our domain of interest, as we could use this raw information available in user-generated text to enrich our existing knowledge about different drugs,diseases and adverse reactions caused by drugs or to monitor adverse events of a specific drug of interest.

We call the problem of linking medical text entities mentioned in informal/colloquial language to a standard medical vocabulary as a *medical concept normalization* or alternatively *medical entity linking* problem. With the term 'medical entity' we refer to a piece of text which represents an entity in the medical domain. For instance, it can represent a disease, an adverse drug reaction or a symptom. The same entity can be expressed in many different ways in written language, for example 'myalgia' and 'myodynia' both refer to the same symptom entity. Therefore, the task of entity linking, tries to identify all the different variations of those entities in text and 'link' them to a standard vocabulary that represents them. This standard vocabulary is referring to the official naming of medical concepts in a knowledge-base. A knowledge base, also referred as ontology in Artificial Intelligence literature, can be considered as a large network of different entities and concepts,where their formal naming and properties, as well as their relations with other entities is stored. In the medical domain for instance, the largest Knowledge-Base named UMLS (Unified Medical Language System) contains structured information about a large number of different medical entities like Diseases, Symptoms, Indications etc, as well as the relations that those different entities have with each other (for instance a certain Disease is associated with certain Symptoms). Being able to normalize entities from user-generated text to their corresponding UMLS entities, would allow us to update the Knowledge Base and hence enrich our knowledge regarding a new adverse drug reaction entity associated with a drug entity, or even update the alternative naming/definition of a certain ADR.

The problem of medical concept normalization is an open problem for the research community. There has already been extensive work done in the field of extracting medical entities like ADRs from users' text, which is a necessary first step, however the nature of the normalization problem, makes it a quite challenging task to solve. Text generated by users in social media is quite different in its linguistics, than using the official medical terminology with which a medical entity is stored in a domain specific knowledge base like UMLS.

user-generated message	**Name of the corresponding KB entity**
'head spinning a little'	Dizziness
'lose 10 lbs'	Body Weight Decreased
'appetite on 10'	Increased Appetite
'terrible headache!!!!'	Headache

Table 1.1: Examples of user-generated text describing ADRs and their related KB entities

As presented in Table 1.1 there are cases with minimal term overlap between the user generated message and the knowledge base entity terms. Hence, understanding the semantics of a medical term mentioned in text, and mapping it to its corresponding entity in a KB is a quite challenging computer science problem, where traditional string matching or term weighting techniques [30] are reported to perform poor [26]. Apart from the difference in the **use of language**, the quite large **number of the available entities** in a KB, as well as the **expensive annotated data** (time consuming process and requirement for domain expertise) from which machine learning approaches could benefit, are also important parameters in this non-trivial problem. In the following sections, we will formalize our research questions, provide an overview of the current state of the art approaches in medical concept normalization and present our novel approach for handling this problem.

1.1 Research objectives and Contributions

In this thesis, our goal is to implement and evaluate a novel approach for normalizing (linking) Adverse Drug Reactions in user-generated text to their corresponding Knowledge Base entities. As a result, we pose the following Research Question.

Main RQ: How can we effectively link medical entities mentioned in user-generated text, to their corresponding entities in an existing medical Knowledge Base?

Of course our goal is not address all medical entities, as the research question poses. Instead we will focus on the specific **case of Adverse Drug Reactions (ADRs)**. In order to answer the above research question, we organize our work around the three research sub-questions as presented below.

RSQ1: What are the state-of-the-art methods, for linking medical entities to knowledge base entities?

To address the above question, our contribution will be to perform a **systematic literature review** in order to identify all the relevant research to medical entity linking. This will let us identify and analyze the current capabilities and limitations of the existing approaches. We will follow a systematic way of retrieving the literature in order to ensure reproducibility, and the exact methodology will be described in detail. The main purpose of this part of our work is to **identify the best performing techniques** in this domain and to **define a research direction that will try to address the limitations** of the current state of the art.

RSQ2: How can we address the drawbacks and limitations of the current state of the art techniques in normalizing ADRs in user-generated text?

Here we will define the whole **solution space** and set up a **novel approach** which, based on theory and related work, can address the limitations of the current techniques and further improve its performance in user-generated text.

RSQ3: How effective is our proposed approach in linking ADR mentions compared to the current state of the art techniques?

Finally, we will perform an extensive **experimental evaluation** of our proposed model and compare its effectiveness to the current state of the art on real world data from social media. As part of this work, we will also evaluate our system in the corresponding shared task of the 2019 Social Media Mining for Health workshop [7], part of the ACL conference. Our evaluation will try to provide the necessary insights about the strengths and weaknesses of the porpoised approach from both quantitative and qualitative perspective.

1.2 Document Structure

The rest of this document is structured as follows. In Section 2 we present the Related Work in the domain of 'medical entity linking' and we perform a Systematic Literature

Review. Based on our findings in this survey, we provide an overview of our proposed approach for normalizing Adverse Drug Reactions (ADR) mentions in user-generated text in Section 3, while in Section 4 we will evaluate it against the state of the art techniques reported in literature. Finally in Section 5 we provide some research directions for future work and present our final conclusions and remarks regarding this thesis.

Chapter 2

Background and Related Work

The goal of this chapter is to inspect and reflect on the evolution of research in the domain of 'medical entity linking'. Our literature survey is based on Systematic Mapping Studies methodology [18, 19], in order to retrieve the most relevant and significant published research in that field. The collected bibliographic data is then analyzed from a perspective that can give an insight to the reader, about the evolution of research in that domain, the main types of techniques that are followed, as well as their strengths and limitations. Our final goal is to extract a useful direction for our research project in normalizing medical entities in user-generated text.

With the term 'medical entity' we refer to a piece of text which represents an entity in the medical domain. For instance , it can represent a disease , an adverse drug reaction or a symptom. The same entity can be expressed in many different ways in written language, for example 'myalgia' and 'myodynia' both refer to the same symptom entity. The task of entity linking, tries to identify all the different variations of those entities mentioned in text and 'link' them with their corresponding entity identifier, as referred in a Knowledge-base.

Knowledge-bases are large network of entities which capture their semantic types, their properties and the relations between different entities. One of the largest knowledge-bases in the medical domain is UMLS, where a very large number of different diseases, symptoms and other medical related entities are stored, capturing any relations or properties between them.

Linking entity mentions in free text , to knowledge base entities is an extremely useful task. It gives us the chance to transform all the unstructured information that is present in text and is related to the examined entity, into structured knowledge that will enrich the existing Knowledge-base information. Building and extending knowledge-graphs (graphical representations of knowledge-bases) is widely used by famous search engines like google in order to improve users' search results. The search engines try to identify the input entities in the search query and retrieve all relevant structured information related to it from a knowledge-base.

In the biomedical domain, there is a lot of useful information in patients' online discussions or doctors' forums, regarding symptoms of a disease or adverse drug reactions of a medicine, which is unstructured and noisy, so it can be lost. As a result, being able to identify those entity mentions in text and link them to their identifiers can en-chance our knowledge by building up an existing knowledge-base.

In the following sections, we will try to present a basic overview of the literature retrieval process, a classification of the different 'medical entity linking ' techniques, the available datasets for evaluating these techniques, a comparison between them and finally demonstrate a useful direction for future attempts in this research field.

2.1 Methodology for literature retrieval

In order to retrieve all the relevant literature on medical entity linking approaches, we followed the systematic literature study approach [18] which is widely adopted in computer science and software engineering. More specifically, we conducted a systematic mapping study based on [19], as it is more appropriate for addressing broader research questions, such as mapping the state-of-the-art methods in a research area. In that way, we make all the implementation steps for retrieving the relevant literature available to the reader, in order to ensure the reproducibility of our efforts. The steps we followed are analyzed in the following paragraph.

The first step was to clearly define our research question. In our case this is, "*what are the state-of-the-art methods in medical entity linking to existing knowledge base?*". From this research question, we then defined the basic keywords related to it. More specifically, the keywords we derived were "*medical*", "*entity*", "*linking*" and "*knowledge-base*". Based on those keywords, our next step was to construct our basic search string which we would use in different search engines. In order to do that we added a set of synonyms to each of the keywords in order to increase the recall of our search results. The overview of our search string can be seen below.

(**medical** OR biomedical OR adr) AND (**entity** OR concept OR text) AND (normalization OR **linking** OR disambiguation) AND (KB OR KG OR (**knowledge** AND **base**) OR (knowledge AND graph))

Having defined our search string, we now started searching for relevant literature in different digital scientific libraries. More specifically, we queried **ACM, IEEE, Scopus and ScienceDirect**. Apart from those digital libraries we retrieved all the relevant documents published in social media mining for health (**SMM4H**) shared task workshop, as one of the published tasks was explicitly related to normalizing social-media health-related posts to official medical terminologies in a medical related knowledge base. Accordingly, we considered the **SemEval e-Health** workshop which also tried to attract researchers that would deal with the medical concept normalization task.

From each of those libraries we collected the relevant documents based on their title and abstract as long as they were published during the last 7-8 years (2012 or later).This resulted in 94 relevant papers retrieved from ACM digital library, 85 from IEEE, 332 from Scopus and 148 from Science Direct. After that initial selection, the retrieved documents of each library were merged together, the duplicates were removed and the papers were filtered based on their title and abstract content. The total number of the retrieved documents after that phase was 86. Finally, we performed a more detailed filtering based on the whole content of each paper which ended up to a collection of 26 relevant documents. Those papers are classified below, based on their common characteristics in order to draw some useful conclusions and directions for our research. A more clear and detailed overview of the bibliographic retrieval procedure can be seen in figure 2.1.

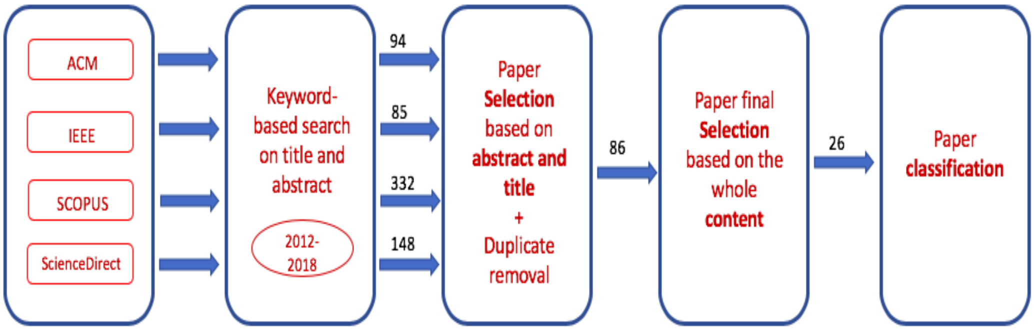
ACM
IEEE
SCOPUS
ScienceDirect
Keyword-based search on title and abstract
2012-2018
94
85
332
148
Paper Selection based on abstract and title + Duplicate removal
86
Paper final Selection based on the whole content
26
Paper classification

Figure 2.1: Literature Survey paper retrieval process

2.2 Classification of different normalization techniques

In this section we will try to provide a broad overview of the different kind of techniques that are used in literature for medical entity linking to an existing knowledge-base. Based on our literature survey results, there are two main categories of medical entity linking techniques: *Rule-based or knowledge-based* techniques and *machine learning* techniques. The second category is quite broad so it can be further categorized to supervised and unsupervised techniques as demonstrated in the tree below. Each of those classes, as well as the most significant related efforts reported in literature are analyzed in the following subsections.

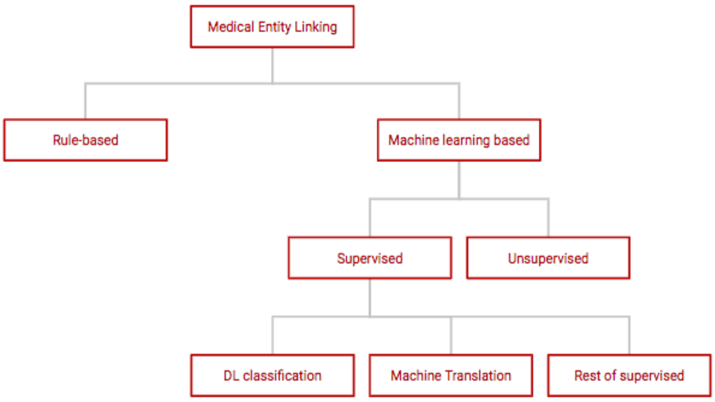

Figure 2.2: medical entity linking categories

2.2.1 Rule-based

Rule- based approaches, sometimes also referred as knowledge-based approaches are handling the medical entity linking problem based on *string similarity measures*, where they are trying to match the text mention with the most similar string representing a knowledge base entity. One of the first attempts was MetaMap [1] in 2001, which we decided to include in our survey despite being a very old approach because it is often used as one of the baselines for comparison with other similar methods. MGrep [43] is another example quite similar to MetaMap reporting a bit better performance, while several others also follow [5, 42].

It is worth mentioning that apart from the traditional string similarity measurements between the text mention and the KB entity, some researchers tried to extend their approaches by implementing dictionary look-up approaches adding mention synonyms or definitions, [11, 6, 28], implementing term weighting techniques [30] or generating candidate entities for a mention using edit distance patterns [11, 17] which is based on the minimum number of operations, deletions or insertions needed to convert a text mention to an entity of the knowledge base.

2.2.2 Machine Learning based

Machine learning based approaches are a quite broad category. For this reason, we have created four main subcategories. The first three of them are supervised machine learning approaches and the last one is non-supervised, in the sense that it does not require any sort of annotated data to be trained with.

Supervised approaches

Deep Learning classification using word embeddings

During recent years, deep learning techniques have demonstrated remarkable results in various tasks including image classification or segmentation but also text classification. Especially Recurrent neural networks have proven to be very useful in tasks related to text as they are able to capture the sequential nature of the text data. Under this assumption, some of the researchers tried to solve the medical entity linking problem as a text classification task, where the text mention is the input to a deep learning model and the corresponding entity of the KB is the label.

The first significant approach was [26], who tried to use a semantic representation of related mentions in user-generated text, as an input to a convolutional neural network (CNN) or a Recurrent neural network (RNN), which applied some filters and extracted features from it in order to correctly classify it to the correct KB entity (class). The above techniques, demonstrated the importance and the usefulness of representing text mentions as vectors that can capture the semantics of this mention. Those vector representations, also called word embeddings, were created using a pre-trained embedding model trained on millions of general domain articles from Gnews.

Inspired from the above work [23] proposed an improvement to the vector representations method. They claimed that creating vector representations of medical concepts based on a general domain corpora does not produce the optimal semantic vectors, and for this reason they trained their word embedding models on a large domain specific unlabeled corpora in combination with the general domain Gnews and they proved that this could give a boost to performance.

More recently [46] proved that the quite promising results presented in [26] were a bit unrealistic because of the fact that the datasets used for evaluation, included the same phrase-label pairs more than once, which could lead to misleading estimations of the true accuracy in previously unseen data. When, removing all duplicates from the dataset, the authors of that work demonstrate results which show that the use of LSTMs and GRU recurrent neural networks are actually performing better than the CNNs reported in previous works in the context of medical concept normalization.

Another family of deep learning methods [14, 35], demonstrate the importance of using character level embeddings with RNNs in order to extract useful features from the entity vector representations. The importance of character-level neural networks relies in the importance of reducing the out of vocabulary (OOV) problem, which is often present in most word embedding models. Especially in social media where spelling mistakes and use of language are totally different than formal text.

Machine Translation approaches

Machine Translation approaches, are also using deep neural networks to link medical entity mentions to their knowledge-base entities, but not as a classification task. Instead they try to 'translate' the 'unofficial' language where the entity is mentioned to the 'formal knowledge-base' language where the KB entities exist. Most of the approaches found in literature try to link entity mentions in user-generated language, so it is mainly a translation from 'user-generated language' to 'knowledge-base language'. For instance, Limsopatham and Collier in 2015 [25], adapt phrase-based MT approach combined with a similarity score between the word vectors to map twitter phrases to SNOMED CT concepts. The aforementioned approach is based on the intuition that entities mentioned in user-generated text are quite different than the official medical terms representing each medical entity in a KB. For this reason, instead of trying to find the similarity between the user text mentions and the labels directly, it is reasonable to embed those different kinds of text in two embedding spaces and then perform some sort of mapping from one space to another, similarly to a language translation approach.

Similarly in [31] the authors utilize an encoder-decoder architecture to translate a disease mention in social media language to a formal medical KB concept. The input to the Bi- directional LSTM encoder is a vector representation of the disease mention, which then is 'unrolled' to KB language, while in [32] authors try to use a quite similar neural network architecture for ICD-10 coding of English Death certificates. In [16], we can find a similar approach to the above, however the deep neural network architecture maps entities mentioned in biomedical literature to KB entities by creating Graph embeddings that capture the relationship between nodes (entities) in the KB.

Rest of the supervised approaches

In this section, we include the most representative supervised machine learning approaches that do not belong to the above two categories (deep learning classification and machine translation methods).

One of the first and probably the most significant supervised machine learning approach was DNorm [21, 22]. In this paper, the authors performed disease entity normalization using a function that performed pair-wise ranking between a disease mention and a candidate concept. This function was trained on given mention-concept pairs to return higher scores when the pairs were correct and lower scores in any other case. Another LTR (Learning to Rank) approach in the field[24], performed CNN-based pair-wise ranking between the text mentions and the KB entity definitions. The first phase of the approach consisted of a candidate generation phase which was rule-based. The authors generated a candidate label list for every entity mention using this rule-based technique, searching for string similarities between the two pieces pf text.

Other machine learning approaches [9, 27, 50], represent a text mention as a set of textual and semantic similarity features and based on them, they train a linear classifier, usually an SVM, to predict the correct corresponding concept.

Unsupervised approaches

When referring to Unsupervised machine learning approaches, we mean approaches that rely only on prior-knowledge which are present in the Knowledge base without making use of any sort of annotated data. Those approaches are using a vector representation of entity mentions that captures their semantic similarity, as well as a similarity measure between the test samples and the Knowledge base entity prototypes. Those vector representations of words and entities are called **word embeddings**. Word embedding models, are trained on a very large unlabeled corpora in order to produce vector representations for any given word present in that corpora. The vectors are created in such a way, that similar words in meaning will have similar vector representations. The model architecture is based on a neural network which tries to predict a word based on its context (CBOW), or to predict a context of a given word (Skip-gram) [33].

As mentioned in previous sections, word embeddings are usually used as an input to deep learning classification methods, for medical entity linking task. However, they are also used to link a word representation of a token , to its most similar vector representation of a KB concept. This similarity is usually measured with the *cosine similarity* between the two vectors. An example of this kind of work is presented in [45] where the authors tried to link anatomical phrases in radiology reports to their corresponding SNOMED CT concepts. The authors experimented with different models for creating the vector representation of a given mention and a SNOMED CT concept and then linked them based on their cosine similarity.

However, we have to mention that to our knowledge there are not many efforts that rely only on unsupervised techniques. There is plenty of research on the other hand, that refers to the word embedding techniques, either as a baseline for comparison with other supervised methods [26, 25] , or calculates the vector similarity as a feature for another supervised classification method [27]. The reason is that, as mentioned in previous sections, the use of language in domains like social media or online forums is totally different than the official medical terminology used in KB and hence it is hard to find a common embedding space that will be able to match semantically similar entities.

2.2.3 General remarks and overview

In this section we will provide a few useful insights by presenting a general overview of the selected research papers in this survey. As can be demonstrated in the pie chart below (Figure 2.3) , the majority of the available research is approaching the medical entity linking problem using machine learning techniques. However, the percentage of the rule-based approaches is larger than any other supervised or unsupervised machine learning subcategory. This can be explained, if we take into consideration that rule-based approaches were probably the only choice before the rise of machine learning techniques and the increased availability of training data during the last few years.

On the other hand, we can also highlight that there is a tendency to switch to deep learning based, and in general supervised learning based, approaches during the last two or three years. As can be seen in Figure 2.4 , rule-based related publications were mostly published before 2017, while on the other hand there is a clear hype in deep learning classification methods and Machine translation approaches. The larger availability of annotated data during the recent years, has clearly played a vital role in this phenomenon.

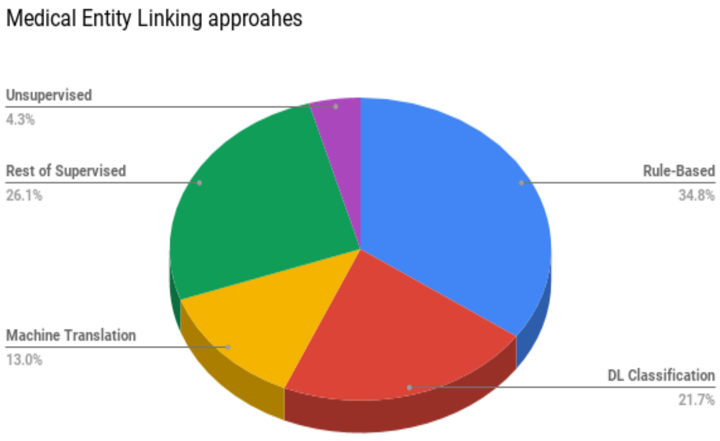

Figure 2.3: Overview of medical entity linking approaches

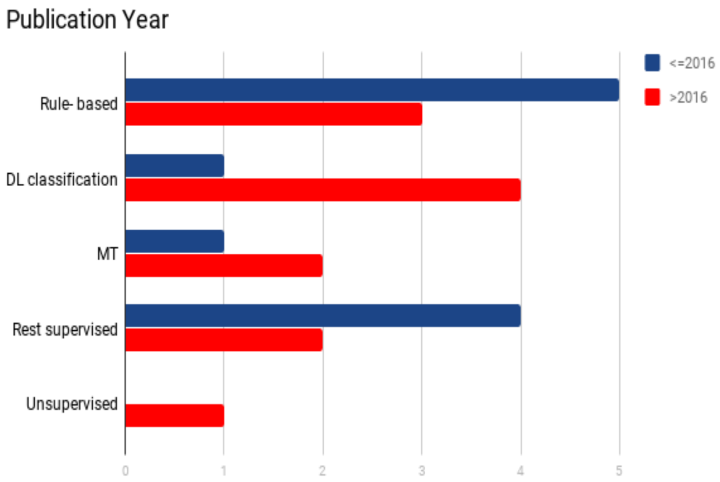

Figure 2.4: Publication year of relevant research

2.3 Dimensions to compare normalization techniques

In this section we assess the medical entity linking techniques presented earlier, with respect to the following aspects of interest:

- Performance in user-generated content.
- Annotation costs.
- Number of hyper-parameters to be tuned.

Based on the above aspects of interest, we are introducing our basic dimensions for comparison of the aforementioned categories.

Performance in user generated-content: The main goal of this literature survey is to identify and present the most significant types of methods used in medical entity linking, in order to give a reasonable research direction for the problem of linking user-generated adverse drug reaction mentions to their corresponding medical concepts in a Knowledge base. For this reason, one of the basic dimensions for comparison is the methods' performance in user-generated content like social media or public forums.

Annotation costs: Since the amount of the available annotated data in this domain is relatively small, and the available number of concept identifiers is quite large, we have to take into account the amount of labeled data that each of the aforementioned techniques needs in order to produce satisfying results in terms of performance.

Number of hyper-parameters to be tuned: Apart from the performance and the need for training data, we evaluate each class of the aforementioned approaches, based on the model complexity. More specifically, we are considering the amount of hyper-parameters that need to be optimized so that each approach can achieve an optimal performance.

2.4 Datasets to study medical entity normalization

Before we demonstrate a comparison between the different medical concept normalization techniques, we will first present all the publicly available datasets for medical concept normalization task in short. The collection of datasets was based on the evaluation sections of the retrieved literature as well as organized workshop tasks in this domain.

2.4.1 Dataset Requirements

The datasets for normalizing medical concepts need to have two basic requirements in order to include them in this section:

- **text entities should be extracted from user-generated text**: As mentioned earlier in this chapter , the scope of this thesis is to normalize medical entities in user-generated text. The use of language in that case, is totally different than a potential official document where the use of language is more official and less noisy.

- **All entities should be mapped to a medical concept from a standard clinical terminology defined in a KB**: This means , in other words, that the annotations for each text mention should be a code from one of the few known medical Knowledge Bases like UMLS or MEDDRA

2.4.2 Publicly available datasets

To our knowledge, at the time of writing this thesis there are 4 publicly available datasets for medical concept normalization in user-generated text. Three of them are extracted from *Twitter* posts and one of them is extracted by *AskAPatient.com* public forum. A brief description of each one of them is demonstrated below.

- **Twitter ADR-S**: This dataset, provided by [25], contains *201* unique *Twitter* phrases describing Adverse Drug Reactions, which are mapped to their corresponding concept from the SNOMED CT[1] clinical vocabulary. The total number of concepts present in the annotations is *58*, which means that each concept has approximately *3.47* text mention examples.

- **Twitter ADR-L**: The TWITTER ADR-L dataset, published in [26], contains 1436 unique TWITTER ADR phrases mapped to 273 medical concepts from the SIDER4 [2] collection of Adverse Events. On average, each medical concept has 5.26 training examples, while 170 out of the 1436 TWITTER mentions are assigned to more than one medical concepts.

- **AskAPatient.com**: The dataset was retrieved from CADEC corpus [15] and consists of reviews from `askapatient.com`.The dataset contains 8,662 phrases, which are annotated to one of the 1,036 medical concepts from SNOMED-CT and AMT (the Australian Medicines Terminology).This means that each medical concept has on average 8.3 training samples. Unlike the aforementioned datasets however, this dataset does not consist of ADR entities only.Diseases (283 entities), Symptoms (275 entities), and Clinical Findings (435 entities) are also added to 6,318 ADR entities.

- **Twitter SMM4H 2017**: In the Social Media Mining for Health Shared Task 3 in 2017, participants were asked to normalize extracted ADR entities from TWITTER posts to their corresponding medical codes from MEDDRA KB. The given training and development set [39], consists of 3629 unique adverse drug reaction mentions mapped to 507 different MEDDRA codes. Unfortunately, the test set annotations are not made public even after the end of the workshop task. However, it is still the largest TWITTER dataset available to our knowledge. On average each MEDDRA code has 7.1 unique training samples.

2.4.3 Selected Dataset for evaluation

In most papers evaluating medical normalization approaches in user-generated content, Twitter ADR-L and AskAPatient datasets are the most commonly used. Despite the fact that Twitter SMM4H 2017 is larger than Twitter ADR-L , it was probably not available when most of the related research took place. For this reason, we performed a comparison of the related work based on their reported performance on Twitter ADR-L and AskAPatient dataset, but then we mainly used the **SMM4H 2017 TWITTER** dataset for evaluating our own work. In order to compare it with the rest, we also reproduced the state of the art approaches from relevant research in this dataset.It is worth mentioning that in order to avoid overfitting,some basic parameter tuning of our models was performed on Twitter ADR-L.

[1] https://www.snomed.org/

[2] http://sideeffects.embl.de/

Our selection was based on the size of the dataset, as it is the largest available in TWITTER language, as well as the nature of social media text which is way more noisy and informal compared to public forums like ASKAPATIENT where patients tend to interact with doctors so the level of noisy,misspelled or slang language is lower.

2.5 Comparing the different normalization techniques

In this section we will present a comparison between the different entity linking techniques, based on the aspects defined in section 2.3. We will try to mention the basic advantages of each one of them, as well as identify their limitations and draw some general conclusions.

2.5.1 Performance in user-generated context

Since every research paper included in this survey evaluates the proposed approach in a large variety of ways, including different datasets and metrics , it was not easy to compare their performance based on a common baseline. However, based on the work presented in[26], as well as in [23, 35, 46], we have the ability to compare the accuracy of the most important representatives of each of the aforementioned categories on two user-generated datasets created from TWITTER posts (Twitter ADR-L) and ASKAPATIENT (AskAPatient.com) forum reviews.

In terms of metrics, since only one 'label' for each text mention is predicted, **accuracy** seems to be the most appropriate one to use. Precision/recall or f-score does not give us any better insights in most cases. However, we have to mention that there are few approaches which perform candidate generation or ranking of candidate concept identifiers, that use precision, recall or a ranking metric like MRR which is more suitable. In our comparison we will stick to accuracy.

From the reported results, we can easily conclude that the **rule-based** approaches demonstrate a relatively poor performance [26] (approximately *0.23* reported accuracy on Twitter data) when they have to do with social media text entity linking. Since they mostly depend on string matching techniques, ignoring the semantics of each text mention in most cases, they seem to fail to link lay people language to official medical terminology. On the contrary, **deep learning supervised techniques** [26, 23, 46, 35] demonstrate a quite promising performance (approximately *0.39-0.46* reported accuracy on Twitter data), as their input is a vector representation of an entity mention that manages to capture the semantics of the phrase and is not dependent on string similarity measures only. The **rest of the supervised** methods, like DNorm, multinomial Logistic Regression [26] as well as **machine translation** approaches [25], are somewhere in the middle (approximately *0.33-0.35* reported accuracy on Twitter data) as they take advantage of the annotated data to learn a function that correctly classifies each mention. Lastly, **unsupervised** techniques, despite the fact that they capture semantic similarity between text with minimal string similarity, they barely outperform the rule-based techniques in those two datasets.

2.5.2 Annotation costs

Regarding the annotation costs, meaning the amount of labeled training data needed for each method, it is obvious that rule-based techniques and unsupervised machine learning techniques have no annotation costs at all. This makes those approaches really easy to acquire. On the other hand, deep learning classification methods and some of the other supervised approaches require a large number of training samples per class in order to perform well, as deep neural networks are very complex models and need a lot of training data to avoid over fitting and generalize well. As reported in [26] the deep learning models clearly overfit when they are trained on a Twitter dataset with approximately 1500 training samples. Some machine translation approaches and other supervised methods that perform pairwise ranking like DNorm, also require a large amount of training data but compared to DL methods they need a large number of entity mention-label pairs and they do not consider different classes.

2.5.3 Number of parameters to tune

The number of hyper-parameters that each approach needs to optimize, plays a significant role in the simplicity or complexity of implementing a model for a specific use case. It is obvious in our case, that deep neural network approaches have the largest number of tunable parameters. Apart from the learning rate, the regularization terms, there are a few other parameters like the dropout rate, the loss function and of course a vast number of trainable weights which makes those kind of models quite complex. Other supervised parameters, like DNorm or Logistic Regression, also need to learn some weights during training but there are less hyper-parameters to tune apart from that. Regarding the unsupervised approach, if we consider the cosine similarity technique using pre-trained word embedding models, then there are no weights or parameters to tune at all. This makes those approaches quite easy and fast to implement. Similarly, rule-based approaches do not have any learnable parameters, but sometimes we need to create vectors like tf-idf of each term etc which makes those approaches a bit more complicated to use that the unsupervised with pre-trained embeddings.

2.5.4 Summary

As a general conclusion from the above comparison, we can say that methods which present the most promising performance are deep learning classification methods. They take advantage of the semantic information captured by word embeddings and using mostly Recurrent Neural Networks with LSTM or GRU layers , they can predict the correct concept label quite accurately. However, they require enormous amount of labeled training data which is the case for every deep neural network. In [46] for instance the accuracy that is demonstrated relies only on concept identifiers that have 5 or more training samples each. So if we take into account that the number of the existing concepts in a knowledge base is huge (at least hundreds of disease symptoms are present in UMLS for instance), then those techniques become too expensive if we want to generalize beyond some very common ADRs that are mentioned quite often in public forums and social media.

On the other hand, rule-based and unsupervised techniques do not need annotated data but they seem to perform poorly especially in cases of user-generated text mentions. This means that we have to find a **trade-off between performance and cost**. On one hand we have to take advantage of the remarkable performance that Neural networks demonstrate but on the other hand we have to take into account some ways to reduce the cost of requiring a very large number of training samples per class to achieve that. In Table 2.1 demonstrated below, we provide a summary of comparison between the different entity linking approaches.

High-Level Approach	**Low-Level Approach**	**Accuracy TwitterADR**	**Accuracy AskAPatient**	**Annotation cost**	**Parameters to optimize**
rule-based	TF-IDF BM25	*0.22-0.23*	*0.55*	No	none
DL classification	GRU LSTM CNN	***0.38-0.46***	***0.79-0.85***	Yes	many
Machine Translation	phrase-based MT	*0.31-0.33*	*0.71-0.72*	Yes	many
Rest of Supervised	DNorm Logistic Regression	*0.3-0.35*	*0.73-0.77*	Yes	few
Unsupervised	cosine similarity	*0.23*	*0.55*	No	none

Table 2.1: Medical Entity Normalization (Linking) approach comparison summary

2.6 Limitations of the state of the art techniques

In the previous section we concluded that the state of the art approach for normalizing medical concept entities in user-generated text, is deep-learning classification using word embedding vectors as input features. In this section, we will try to further analyze the aforementioned limitations of those approaches, which we will try to fill with our research described in the following chapters.

2.6.1 Limitations

Despite their high performance demonstrated in different datasets of user-generated text, we have already mentioned that is a quite expensive approach in terms of the amount of annotated data that it needs. In fact, it is the nature of the medical concept normalization problem that introduces those limitations to this approach. There are two main dimensions of the problem that pose this.

First of all, some medical concepts for instance some ADRs or symptoms, are way more common than others is social media of public forum posts. It is common to find many different expressions referring to ’Headache’ or ’Stomach Pain’ caused by a drug use, rather than ’sleepwalking’. As a result, there is a high class imbalance issue between some common medical concepts that usually show up in social media and some others that are quite rare. To demonstrate this, we have added a pie chart from the SMM4H 2017 Twitter dataset shown in Figure 2.5, where we can clearly see that more than 40 % of the medical concepts present in this dataset (507) have just one training sample.

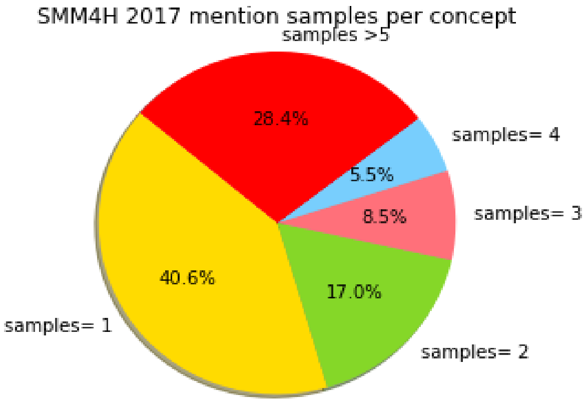

Figure 2.5: Available training samples per concept in SMM4H 2017 Twitter Dataset

Deep neural networks in general, need a large number of training samples in order to demonstrate a remarkable performance. Otherwise, they have a tendency to overfit. Current deep neural network approaches, handle all medical concepts in the same way, as they consider each one of them as a separate class. So our hypothesis, is that RNNs and CNNs demonstrate a good performance, because they are able to successfully normalize some common adrs in nature which compose a significant part of most datasets.

However, it is worth analyzing their accuracy across all different concepts that they are trained to predict. In Figure 2.6 below, we have reproduced the state of the art RNN as presented in [26], to visualize the performance of this model as a function of the available training samples that each concept (class) has.

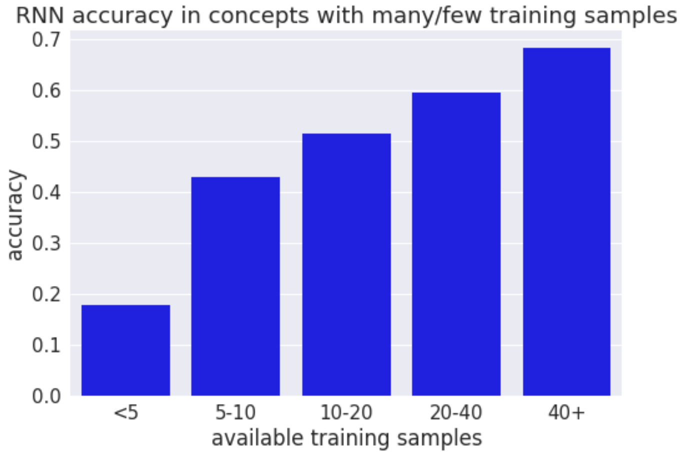

Figure 2.6: RNN with single GRU layer performance as a function of the available training samples

The second problem that we have to consider, is that there is not a fixed number of medical concepts that are available and between which we have to select. Depending on the KB, there are thousands of symptoms or ADRs stored and it is possible that during test time, we will have to normalize a concept that was never seen before as it does not belong to one of the known medical concepts present in our retrieved training data. This is an extension of the first problem we mentioned in the previous paragraph , however it makes the successful normalization of those concepts impossible with a supervised approach that is trained on a specific training set only.

To summarize,the above means that deep neural networks are quite good in demonstrating a remarkable performance in some small datasets of user-generated text, but in real-world the nature of the problem is totally different. Our research, should focus on a less sensitive to the presence of training data approach, while at the same time it will be able to demonstrate a competitive performance in some common medical concepts, where the variation of the available data allows us to use complex supervised models to learn useful features from them. Our proposed approach as well as its evaluation are extensively presented in the following chapters.

Chapter 3

A Novel Few-Shot Learning Approach for ADR Normalization

The goal of this chapter is to provide the reader with a general overview of our selected approach for normalizing medical concepts in user-generated text. Our approach aims to fill in the gaps and limitations identified in the current state of the art techniques, as presented in the previous chapter of this document. In the following subsections we analyze our reasoning and motivation behind the selected approach, a high-level description of our model as well as a more low-level description of the model parameters, selected during our implementation phase.

3.1 Discussion of current Approaches

As mentioned in the previous chapter there are two main limitations when handling the medical concept normalization problem as a standard deep learning classification task.

- many concepts have limited and insufficient training examples available
- some previously unseen concepts may appear at test time

Based on the above, it becomes obvious that we have to find an alternative way of predicting concepts when we have insufficient or even zero training samples. In the medical domain, getting more annotated data for those rare concepts would be extremely expensive and time consuming. On the other hand, creating synthetic samples would seem like a reasonable solution to boost the performance of the state of the art neural networks. However, the number of classes that would need a significant number of synthetic training samples is so large that it could lead to a training data set, where the percentage of the actual user-generated examples would be a minority. Then how can we predict a class (medical concept) for which we have 1 or maybe 2 representatives? In machine learning, this is a typical **few-shot learning** problem case. Few-shot learning [47], refers to machine learning algorithms that are able to perform a prediction using only few 'shots' from each class, contrary to normal practise of using a large variety of training data in deep learning or other ML models. In the following section, a more extensive analysis of few-shot learning algorithms and how they are used is presented.

3.2 FSL Background

Few-shot learning (FSL) approaches are widely used in the computer vision domain for image classification, image retrieval or image tracking. A typical example of FSL is handwritten character recognition, where the computer classifies handwritten characters based on their similarity with a few given examples of those characters. FSL is usually used when there is a need to reduce the extra data gathering effort or when we need a model to learn from rare cases with limited annotated data. It can focus either on data augmentation techniques on an existing traditional ML approach or on different models which are capable of learning from few 'shots' (examples). An overview of the different families of models used in few-shot learning, as categorized in [47], are demonstrated below:

- *Multi-task Learning methods*: Those methods take advantage of the similarity between different tasks to transfer parameters of a model, which is trained on one task, to another similar task with fewer training data.
- *Embedding Learning methods*: Embedding learning methods, embed the input sample into a smaller embedding space where similar from dissimilar objects can be easily identified. It is quite common approach in literature and is mainly used for classification tasks. There are three key components for an embedding learning approach: The embedding function $f(.)$, which embeds the samples of the support set (training set) to a different space, the embedding function $g(.)$ which embeds the unseen test samples to the same space as the training data and finally the similarity measure s(.) to compare the embeddings of the training and test samples.
- *Learning with External memory*: Neural Turing Machines [13] and memory networks are typical examples of Learning with external memory approaches. In these approaches when a new task is given, instead of training our model again which is costly,they directly memorize needed knowledge in external memory. Then similarly to the embedding learning case, a new unseen sample is embedded using an embedding function f and then it is compared with each embedded object stored in the external memory.
- *Generative Modeling methods*: This family of FSL methods use the available prior knowledge to learn a probability distribution. Then the learned generative models can generate artificial samples as an augmentation technique. Typical examples of Generative models are the Variational Autoencoders (VAE) and GANs.

In *Natural Language Processing* and specifically in text classification, there is no extensive research available in few-shot learning approaches, compared to computer vision. However recent developments in distributional semantics [33, 37, 3], have demonstrated a very useful unsupervised approach to represent terms as vectors in a semantic space (word embeddings).

As our FSL approach, as well as the current state of the art approaches in medical concept normalization, are based on the vector representation of words we will provide the reader with an overview of the basic algorithms we use to create those embeddings. The first word embedding algorithm named word2vec [33], is based on a simple neural network architecture with a single hidden layer , which is trained on a large corpus of unlabeled text. Word2vec receives the context of a word as input and predicts a target word. As it is not possible to feed a word as a string into a deep neural network, the input words are fed as one-hot vectors , whose length is equal to the number of the unique words in the training corpus, and which are filled with zeros except at the index that represents the word we want to represent, which has a value of 1. The output of the network is a softmax activation function which returns a probability for each one of the target words that are present in the vocabulary. The network is trained to predict the correct target words for each input text and finally the weights of the hidden layer are the word embeddings for each one of the one-hot encoded words. In other words,

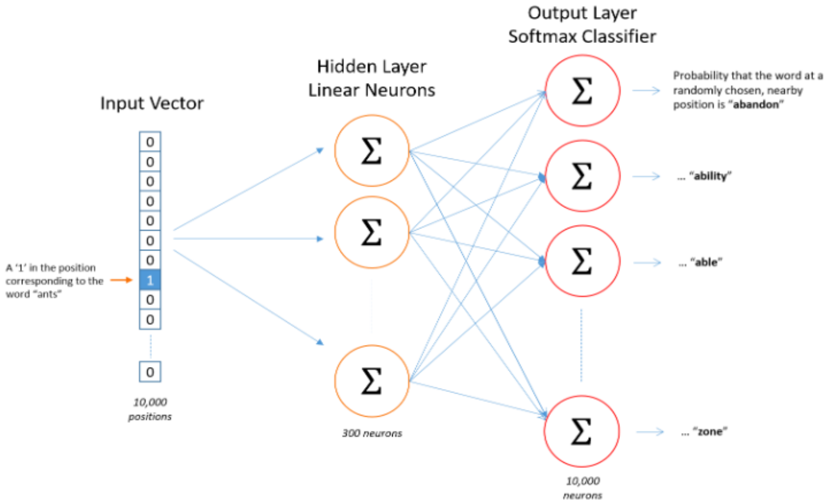

Figure 3.1: word2vec Algorithm

after the network is trained the hidden layer acts like a look-up table for the embeddings of each word as demonstrated in Figure 3.2. Technically, word2vec is motivated to learn similar vectors (embeddings) for words with similar context, which makes it capable of capturing semantic and syntactic properties of the different words no matter what their degree of string similarity is.

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Figure 3.2: Word2vec Hidden Layer

The above research has inspired the most representative attempts of FSL in this field, which try to take advantage of those term representations in order to create an intermediate fixed-size *vector representation on the phrase/sentence level* which is able to capture its semantics. Word embeddings are capable of producing high quality vector representations on the word level, however those approaches aim to use those representations in order to represent multiple length phrases with a fixed size vector too, in such a way that the semantics of the whole phrase are preserved. Then using a similarity metric or a classifier on top of those representations, they classify the unknown samples based on the prior knowledge they have from the support set. In short, we can categorise the current approaches on few-shot text classification as follows:

- *Siamese Neural Networks*: Siamese Neural networks are deep learning architectures trained with pairs of similar or dissimilar text. The multi-token input phrases are usually embedded using a pre-trained word embedding model and then the siamese network is trained to output fixed-size vectors for each phrase of the input pair [48]. Then network is trained to minimize the Manhattan distance between those fixed size output vectors.

- *Encoder networks*: Encoders are also neural networks architectures used to extract useful features and create a fixed size vector representation for different length input sequences. The difference between siamese NN and encoders is that encoders are not trained in pairs of similar or dissimilar text. The input phrase is transformed to a fixed dimensional embedding space which is then followed by a classifier [49].

- *Individual Token embedding aggregations to a fixed size vector*: Instead of using neural networks to produce fixed size representations, some approaches take a more simple approach to achieve that. Especially when the size of the text is not large, averaging, adding or performing some sort of hierarchical pooling of the different token embeddings of a phrase , are capable of creating a meaningful fixed-size representations of a multi token input text, in case the length of the text is short enough. Then as a final step, this vector is either fed into a classifier [36] or a simple distance metric between those vectors is used[2].

3.3 Introducing our approach

Our selected approach, taking into account similar research in the FSL domain and the nature of our problem, is demonstrated in Figure 3.3 on a high level. Our approach is more close to [2], in the sense that our embedding function $g(.)$ for creating a vector representation of the input phrases, is based on a simple aggregation of the individual token vectors of each ADR. The token vectors are retrieved from a pre-trained word embedding model. For multi-token ADR phrases, the pre-trained word embedding model is used to embed each token into a d-dimensional vector. Then a simple aggregation (i.e element-wise addition, average) of the multiple d-dimensional vectors of each token will end up in a d-dimensional vector representing the whole phrase. After the projection of every phrase from the train and test data into this d-dimensional vector space, the normalization/classification of the test samples is done based on their nearest neighbor from the training data. The exact function $g(.)$ as well as the optimal similarity metric s will be determined in the next section of this report. There are

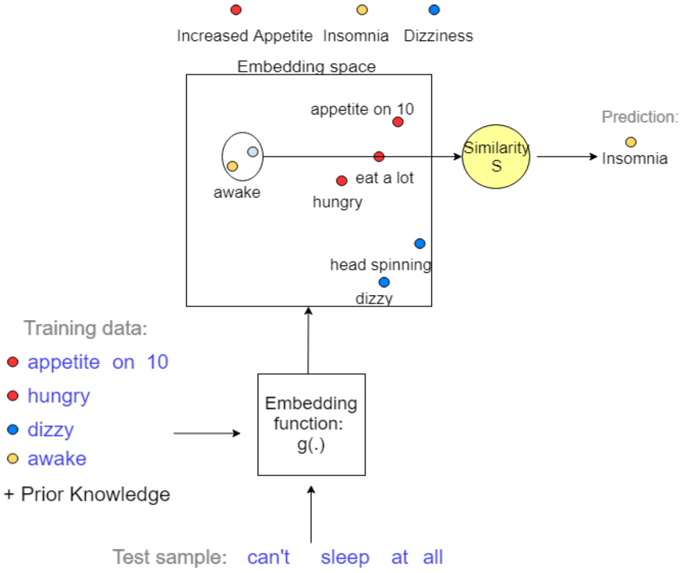

Figure 3.3: Approach Overview

three basic reasons that led us selecting this approach instead of a Siamese or Encoder architecture:

1. *It is a parameter free approach*: Since we want to focus on improving the accuracy of normalization in medical concepts that are rarely seen at test time, we have to be very careful in order to to avoid overfitting. Using a Siamese neural network or any other complex supervised approach with a large number of training parameters to tune, would be a risk. Especially in our case, where the amount of annotated data is limited. On the other hand, SWEMs [40] (Simple Word Embedding Models) which perform simple pooling operations in the different token embeddings and then feed them as input to simple linear models, have been proven to demonstrate comparable or even superior performance compared to deep neural networks in several text classification tasks. [40].

2. *ADR mentions are in their majority short pieces of text*: In most cases, users in social media tend to express their feelings using a few words. We will rarely find a case where a patient describes and Adverse Drug Reaction in a whole sentence. To demonstrate this, we have grouped the ADR mentions of the 3 largest available user-generated datasets in figure 3.4, based on their number of tokens. This fact plays an important role in our approach selection. In other few shot learning problem cases, for instance in Question Answering problems, the corresponding phrases may be one or more sentences or even a whole paragraph. In that case it would be impossible to create a meaningful fixed size representation using simple pooling and aggregation techniques as the amount of noise would be large. So in that case, the use of a more complex network architecture like a Siamese CNN would be a more reasonable choice. In our case, as can be seen from our data distribution the majority of ADR mentions are just a single or 2-3 tokens.

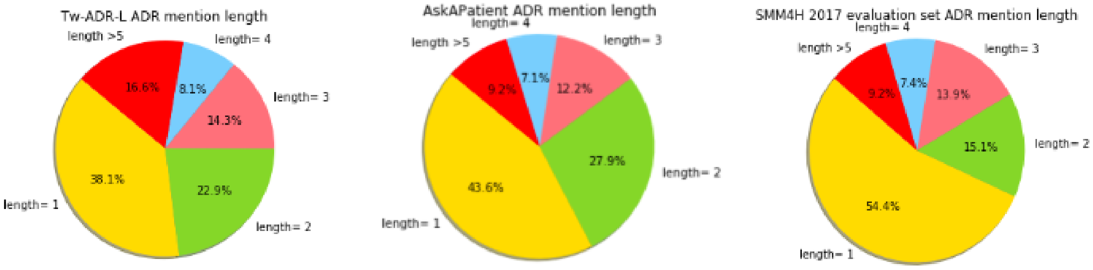

Figure 3.4: Mention Length in different ADR datasets

3. *The use of Siamese Neural Networks*. As an initial attempt in our work, we tried to handle the medical concept normalization problem as a ranking task using Siamese Neural networks. This is a typical approach for entity linking in several domains, where for each query (Twitter mention) a set of possible candidates is retrieved (entities from KB) and then a ranking system like a Siamese neural network is used to predict the highest similarity between the retrieved candidates and the query. While experimenting with this approach we came across two major problems. First of all, in our domain the candidate retrieval process was very complicated and resulted in a very low recall. In most cases, the candidate retrieval step is based on a set on rules related to the string similarity or term matching level between a query and a candidate entity. However in the medical domain, trying to retrieve candidates for user generated twitter mentions was a very difficult task because of the large gap between informal Twitter language and official medical terminology. Secondly, the training process of a siamese neural network is not straightforward. As described earlier, Siamese neural networks are trained on pairs of similar and dissimilar text. The network output is a probability that the text pair is semantically same or not. Therefore, we needed pairs of similar and dissimilar ADRs to train our network with. As the neural network tries to extract features from both parts of a text pair, the number of trainable parameters of the network increases compared to simply classifying one input ADR mention. This requires a large mount of training data to avoid overfitting. Apart from that, creating pairs of positive (similar) as well as negative (dissimilar) text was not a trivial task given the format of our annotated dataset. In order to create pairs of positive text, we needed to connect ADR phrases that were annotated to the same concept (class). However this did not always result in text pairs that were appropriate training samples for our neural network. For instance, the phrase '*awake for 30 hours*' and the phrase '*like a zombie all night*' both refer to '*Insomnia*' but they are not very similar phrases semantically. Creating pairs of negative text was also difficult as some classes, like '*Hunger*' and '*Increased Appetite*', were close to each other and therefore some of their training examples did not have large semantic difference. Based on the above, as well as the fact that it was difficult to determine the optimal architecture of the siamese network for our case, we decided to abandon the above technique.

On a theoretical level, our basic hypothesis is to test whether creating fixed-size vector representations of ADRs from the training and test data is a valuable feature in order to normalize the unlabeled ADR mentions, based only on their vector similarities. In that case, we are expecting that a few-shot learning approach can take advantage of its simplicity and demonstrate a better performance in medical concepts where training data is limited. As we mentioned before, this is a significant percentage of the data in a real world scenario. Hence , we would expect that the overall performance also improves. Our second hypothesis, is that **prior knowledge** from a medical Knowledge base can be a valuable source of data for representing medical concepts that are previously not present in the training data. In every Knowledge base related to the biomedical domain like UMLS or MEDDRA for instance, each entity is assigned with one or more definitions and terms describing that entity. So even if we have no training samples for a medical concept, we could use all different synonymous terms associated with this concept as its representatives, or even enrich the representatives of a concept that has limited user-generated training data. It is important to mention here that prior knowledge could also be used as additional training data to a neural network model, however this choice includes a serious risk. The official medical terminology in knowledge base terms, is not representative in many cases of the user-generated text that the model will have to classify at test time. This can lead us to train a model for a task that is quite different at test time compared to the training process. Using prior knowledge to the aforementioned FSL technique also adds some amount of noise of course, but since no training process is required we expect the impact to be minimum. Finally, compared to the state of the art neural networks, we are assuming that a FSL approach will be less competitive as the number of the available training samples increases. However, it remains to be seen to what extend this will happen and if our novel FSL approach is an overall best solution, if it is an alternative which can be used in a combination with the current state-of the art deep learning approaches, or if the deep neural network alone still achieves a better performance in all the aforementioned subregions of the problem.

3.4 Implementation

After introducing our approach on a high level as well as our motivation behind it, we now need to fine-tune our model parameters. The main parameters we need to select, as demonstrated in Figure 3.3 of this report, are the *embedding function g(.)* as well as the *similarity measure s* between the embedded vectors of the ADRs. For the similarity measure, since we measure distance between two embedded text vectors, we have two main choices. Either *cosine similarity* or *Word Mover's Distance* [20].

Cosine similarity is the most common similarity measure between word vector representations. It calculates the cosine of the angle between the two vectors A and B.

$$cosineSimilarity = cos(\theta) = \frac{AB}{|A||B|}$$

Since similar words or phrases have similar vectors, their cosine distance also indicates semantic distance. It is preferred against simple Euclidean distance which is related to the vectors' magnitude, as it is common [29] that the word embedding vector magnitude is highly correlated with the number of occurrences of one word in the corpus. Since two words/phrases may have the same semantics but one may appear more often that the other (in the corpus used to train the embedding model), the angle between the vectors is a more appropriate metric than the magnitude. In addition, in higher dimensions the cosine similarity range is still between -1 and 1, while the Euclidean distance between the vectors is usually very large.

On the other hand, another popular metric to measure similarity between phrases is *Word Mover's Distance*. The WMD calculates the dissimilarity between two phrases as the minimum amount of distance that the embedded words of the first phrase need to travel in order reach the embedded words of the second phrase. This approach is quite straightforward and requires no extra learning parameters. A simple example of the WMD between two sentences is demonstrated in Figure 3.5.

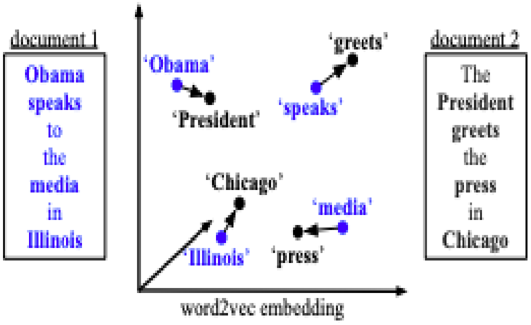

Figure 3.5: Illustration of the WMD approach for phrase similarity

Based on the above, our *embedding function g(.)*, heavily relies on the distance metric that we will use. If we use WMD, then the embedding function to use is straightforward. We only need to select a word embedding model to transform all tokens of our ADRs to vectors and then the WMD between the two ADR phrases can be calculated. If however we use cosine distance, then we have to create a fixed size vector representation of our multi-token ADR text mentions. Word embedding models will give us a vector representation for each individual token, so we need to find the optimal way of aggregating those individual embeddings. Based on [40], we will try *four different aggregation techniques* for creating a fixed size vector representation of each ADR, in order to compare each of the created vectors using cosine similarity.

Summarizing, we will try to find the optimal version of our Few-shot learning approach, experimenting between the following 4 set-ups.

- *element-wise addition of individual token vectors + cosine similarity*: In this set-up , we add the individual token vector representations of an ADR to get a fixed size vector representation. Finally, the most similar ADR to a new unlabeled sample is measured with the cosine similarity between the rest of the labeled vectors.

- *average of individual token vectors + cosine similarity*: Similarly to the above set-up, we now try to average the individual token vectors to create a fixed-size vector representation

- *weighted average of individual token vectors + cosine similarity*: In order to give larger weights to words that are more important than others we calculate the weighted average of the token embeddings. The weights are determined by the tf-idf value of each token in the training data.

- *element-wise max pooling of individual token vectors + cosine similarity*: This set-up takes the maximum value along each dimension of the individual token vectors

- *embedding individual tokens + WMD*: As mentioned above, in this set-up we do not need to aggregate word embeddings into a fixed size vector.We only need to find the minimum amount of Euclidean distance that the embedded words of the each ADR phrase need to travel to reach the embedded words of another ADR phrase.

In order to determine the most effective variation of our FSL approach we performed the following experiment on the SMM4H 2019 Twitter training dataset. The dataset consists of 1212 ADR text mentions from Twitter, normalized to 319 medical codes from MEDDRA Knowledge-base. More information about this dataset can be found in section 4.2

1. We randomly split the SMM4H 2019 dataset into training set and development set (90-10 split).

2. We converted all ADR mentions to lowercase.

3. We removed non alphanumeric characters like (#,! etc)

4. We removed all stopwords from ADRs

5. We created a vector representation of all ADRs in the training data using one of the 5 different aforementioned techniques

6. We created a vector representation of each ’unlabeled’ ADR in the development set.

7. Classified the unlabeled ADR based on its closest vector from the training data, using cosine distance or WMD.

8. Evaluated each one of our models based on classification Accuracy: how many correctly normalized ADR mentions were achieved as a percentage of the total development set.

In this experiment we used a pre-trained 300 dimensional word embedding model, trained on a Google news corpus of 100 billion words with word2vec algorithm [33]. It contains word vectors for more than 3 million English words and to our knowledge it is the largest publicly available pre-trained word embedding model. For the stopword removal step, we used the NLTK Python library, which includes a list of common used stopwords in English language. As we can see, the accuracy of the different aggregation techniques demonstrated in Figure 3.6, is higher when we add up the individual word embedding vectors and then use the cosine distance to compare them. Averaging word vectors also achieves similar performance, while the worst performing aggregation technique is the weighted-average of the word vectors.However, all approaches are relatively close.

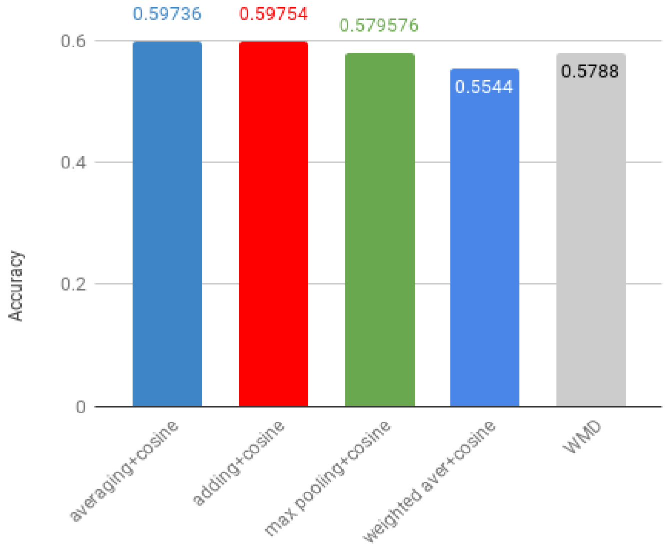

Figure 3.6: Accuracy of different versions of our FSL approach in SMM4H 2019 development set.

Taking into account that adding the individual word embeddings of an ADR results in the most effective vector representation for the whole phrase, we also wanted to test how different word embedding models can affect our FSL performance. As already discussed in previous sections, word embeddings are created using a large corpus of unlabeled text, where each word's context plays the most important role in its representation. Since we have to do with user generated text, and more specifically with Twitter, it would be worth trying to create the embedding vectors of each word based on Twitter corpus as the informal/slang nature of social-media language can affect the meaning of some words and generate representations for domain specific words that are not present in more formal articles from Google News. For this reason we experimented with two different models trained on Twitter corpus. One is trained with GloVe algorithm [37] on 2 billion Tweets and contains 1.2 million word embeddings and the other one is trained with word2vec on 400 million tweets containing a vocabulary of approximately 3 million words [12]. To compare the different models we used our best performing aggregation technique, which is adding the individual token embeddings.

As can be seen from Table 3.1, the general corpus model trained on Gnews achieves a superior performance compared to both domain specific models trained on Twitter corpus. However, it is worth mentioning that Twitter models have fewer out of vocab-

Corpus	**Algorithm**	**Dimensions**	OOV	**Accuracy**
Google News	word2vec	300	102	0.597
Twitter	word2vec	400	49	0.589
Twitter	Glove	200	68	0.586

Table 3.1: Different word embedding models performance in SMM4H 2019 dev. set

ulary (OOV) words , compared to the general domain model which is expected.

To conclude, based on the results of the above experiments our best performing parameters for the Few-shot Learning approach are the following:

- *embedding function g(.)*: As an embedding function of the ADR phrases in the training set and the test set, we use the addition of the individual word embeddings of the Adverse Drug reaction. The word embedding model is pre-trained on a corpus of 100 billion words from Google news, using word2vec algorithm.
- *similarity s*: As a similarity measure we are using the cosine similarity between the embedded ADR vectors.

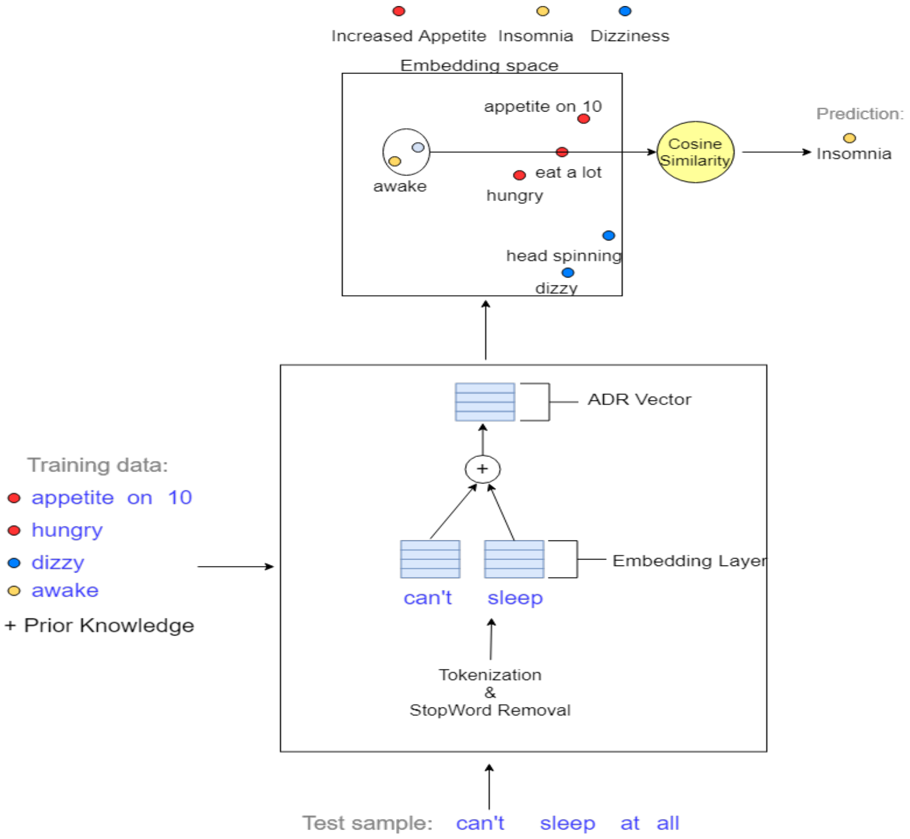

Figure 3.7: FSL Approach Low-Level Overview.

Chapter 4

Experimental Evaluation

In this chapter we will perform an extensive experimental evaluation of our proposed approach on real world data from Twitter. In the first subsection we will briefly present the main hypothesis on which our FSL system is based and then we will perform a set of experiments comparing the proposed FSL technique to the state of the art methods from relevant research. Based on this evaluation we will accept or reject each one of our aforementioned hypothesis and conclude on the feasibility of our few shot learning technique in the medical concept normalization domain. Furthermore, we will demonstrate the results of a qualitative analysis, as well as some additional quantitative findings that will help us identify in more detail the strengths and the limitations of the proposed approach.

4.1 Approach Hypothesis

The main research question that we will try to address in this chapter is '**How effective is our proposed approach compared to the current state of the art techniques**' as posed in the Introduction section of this document. Our proposed few-shot learning approach is based on the hypothesis that it will be able to normalize medical concepts without the need of extensive amount of training samples per class as opposed to the state of the art deep learning models. Therefore in order to argue on the effectiveness of our proposed model we will need to confirm or reject the following hypothesis on which our work is based on:

H1: Our Few-Shot Learning Approach will perform better than the state of the art in normalizing medical concepts with limited training samples. The state of the art Neural Networks will outperform the few-shot learning approach on classes (concept) with a large availability of training samples.

H2: We can combine the proposed FSL approach with the current SOTA deep neural networks to achieve a more robust performance among the different imbalanced classes.

H3: Prior knowledge from a medical ontology can be used as an alternative source of data to normalize medical concepts when no training data is available.

In the following subsections we will try to confirm or reject the aforementioned hypothesis in order to give a complete answer to our posed research question. Before presenting our experiments, we first provide the reader with a general description of the datasets used for evaluation.

4.2 Dataset Description

As mentioned in the Related Work section , we will evaluate our approach on the largest available Twitter dataset to our knowledge. **Twitter SMM4H 2017** was published as part of the Social Media Mining for Health workshop in 2017. Unfortunately , the annotation data was only released for the training set and the development set. The annotations for the test set were not released in public. For this reason , we use the annotated part of the data for our experiments. The development set and train set were originally concatenated and then split into **5 equal folds**. However , out of the approximately 9500 mentions only **3629 ADR mentions** were unique. Those mentions were mapped to **507 medical concepts** from the MEDDRA Knowledge-Base. An example of the dataset is demonstrated below. Each MEDDRA code corresponds to a medical term from the KB vocabulary. For instance, *10020765* is mapped to term '*Hypersomnia*', while *10061428* is mapped to '*Decreased Appetite*'.

ADR mention	**MEDDRA Code**
sleep for 15+ hours	10020765
didn't eat a thing	10061428
withdrawals	10048010

Table 4.1: SMM4H 2017 dataset example

As there was a very high percentage of overlap between the training and test folds, we decided to remove all duplicates in order to avoid over optimistic estimations of our performance. After the duplicate removal, from the validation (development) and test set folds, the final test sets consisted of *approximately 400 ADR mentions for testing and 3200 mentions for training in each one of the 5 different folds*.

In addition to evaluating our approach on the aforementioned dataset, we formed a team which took part in the **2019 SMM4H Workshop shared task 3**, part of the ACL 2019 conference [7]. More information about Task 3 of this workshop, entitled 'Normalization of ADR mentions in social media', will be included in the corresponding section. The given training data for this task consisted of just 1212 unique Twitter ADR mentions mapped to 319 MEDDRA codes. This dataset was randomly split into 90 % of the data used for training and 10% of the remaining data used as development set to tune the hyperparameters of our model and tune the parameters of the SOTA techniques we used for comparison. The test data for this task consisted of 100 Tweet IDs which contained approximately 500 ADR mentions.

Dataset	# ADR mentions	# MEDDRA Codes
SMM4H 2017	3629	507
SMM4H 2019	1212	319

Table 4.2: Summary of Datasets used in the experimental procedure

4.3 Hypothesis 1 Testing: Evaluating FSL Performance

Having already selected the optimal parameters for our few- shot learning approach, we are evaluating it on real world data in order to test if our hypothesis is true. The effectiveness of our few-shot Learning approach was tested against the current state of the art deep learning techniques in SMM4H 2017 Twitter dataset , as well as against other researchers that participated in the 4th Social Media Mining for Health 2019 Workshop (SMM4H2019), part of the ACL conference for computational linguistics.

4.3.1 Reproducing the State of the Art

In order to enable a direct comparison of our FSL approach, we reproduced two state of the art neural networks. The first one is a Convolutional Neural Network and the second is a Recurrent Neural Network with a single GRU layer. Those two models, published in [26], demonstrated superior performance compared to all other rule-based or ML based techniques and besides that the authors made their implementation details public to the research community in order to ensure the accurate reproducibility of their approach. The CNN is composed of a single convolutional and a pooling layer as demonstrated in Figure 4.1. The input ADR text mention is represented as a sequence of d-dimensional word embeddings, where each vector is derived from a pretrained word embedding model. Then a convolution operation is applied to the representation of the input ADR , using a filter w with a window of h words. The output of the convolution layer forms a fully connected layer, which is then passed through a softmax classifier for multi-class classification.

Similar to the CNN, the RNN model (Figure 4.2) deploys a GRU recurrent layer in order to better capture the sequential nature of the text. The output of this Recurrent layer is passed as input to a softmax layer for the classification task.

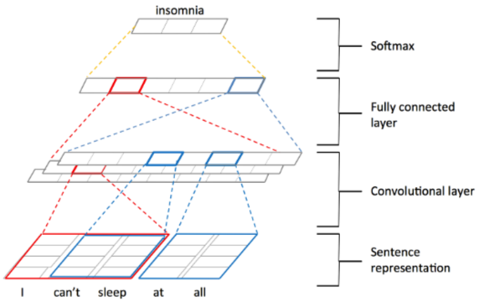

Figure 4.1: State of the art CNN architecture

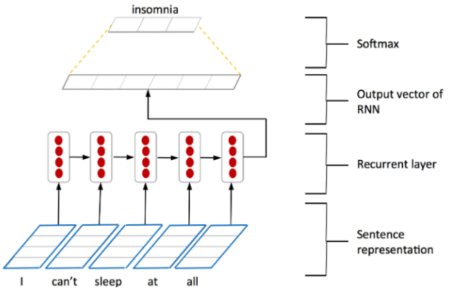

Figure 4.2: State of the art RNN architecture

4.3.2 Evaluating FSL Approach on the Fourth Social Media Mining for Health (SMM4H) Shared Task at ACL 2019

Our team entitled MYTOMORROWS-TU DELFT participated in subtask 3 of the 2019 Social Media Mining for Health Applications (SMM4H) [7] workshop, which is an end-to-end task. The goal of this task is, given a tweet, to 1) automatically classify tweets containing an adverse drug reaction mention; 2) extract the exact ADR mention; 3) normalize the extracted ADR to its corresponding Medical Dictionary for Regulatory Activities (MEDDRA) code.

Since the topic of this thesis is the normalization of already extracted ADR mentions from social media, our contributions focus on the normalization step and linking ADRs to their corresponding MEDDRA code. However, to be able to perform an end-to-end evaluation, we used existing state-of-the art techniques for subtask 1 [38] and 2 [4], which we trained on the workshop dataset. Since the task was end-to-end and the exact spans of each ADR entity were not provided, the evaluation was based on strict and relaxed F-score, precision and recall. Of course the nature of the task did not allow us to perform direct comparison between our normalization approach and the other participating normalization systems, however we were able to keep the first two steps identical, and then evaluate our approach against the state of the art NN on the normalization step only.

For each subtask of this shared task, the participants were provided with an appropriate training set. For more information please refer to [7]. For the normalization subtask, all participants were provided with a dataset of 1212 ADR twitter mentions which were mapped to 319 MEDDRA codes as mentioned in the previous section. At test time, all teams were provided with approximately 1000 Tweets, half of which contained an ADR mention.

For evaluating our approach we initially split our given training data to a training and a development set randomly as mentioned in earlier sections.The results of our best performing FSL approach against the state of the art RNN and CNN can be seen in Table 4.3.[1] As can be seen the best performing Few Shot Learning approach seems to perform better in general than the state-of the art approaches in our development set. However, we have to take into account that there is a high risk that this result is overoptimistic. First of all because, we used the development set to optimize the parameters of the few-shot learning approach and secondly because the development data samples are limited. However, it is worth analyzing this performance to get an insight

Approach	Dev. Set Accuracy
FSL	0.597
RNN	0.571
CNN	0.5

Table 4.3: Development Set performance of FSL and SOT techniques

[1] The RNN and CNN results are the average of 5 different runs on the training data.

about the effectiveness of the few-shot learning model in predicting rare concepts. For this reason, we plotted the performance of the RNN, which performed much better than the CNN in this context, and our approach as a function of the available training samples that each medical concept has in the training data. As demonstrated in Figure 4.3, our initial hypothesis that a simple few shot learning approach with few learnable parameters would perform better when the available training samples for a concept are limited, is confirmed in that case. On the other hand, as the training samples per concept increase the RNN seems to take advantage as it can extract more useful features than just the cosine distance of a vector representation. Of course the difference in this case seems marginal. As we mentioned earlier those results can be a bit misleading, so evaluating those two best performing approaches on the test data will indicate if those results can generalize in new unseen data. Since the task was end to end, and we needed to extract the correct ADR spans before normalizing, we run the first two subtasks using the state of the art methods once, and used their output to normalize the extracted ADRs with the FSL approach and the RNN neural network. The average relaxed and strict F score of our two systems, are demonstrated in Table 4.4. As can be seen, despite the fact that we can not calculate the exact accuracy of the normalization subtask, the total performance of our FSL system is still better than the same system (we ony run steps 1 and 2 once and used the results for FSL and RNN) which uses the RNN for normalization. Apart from that, our approach seems to be above the average of the other systems that participated in this task, as we ranked second among the 4 teams that participated.

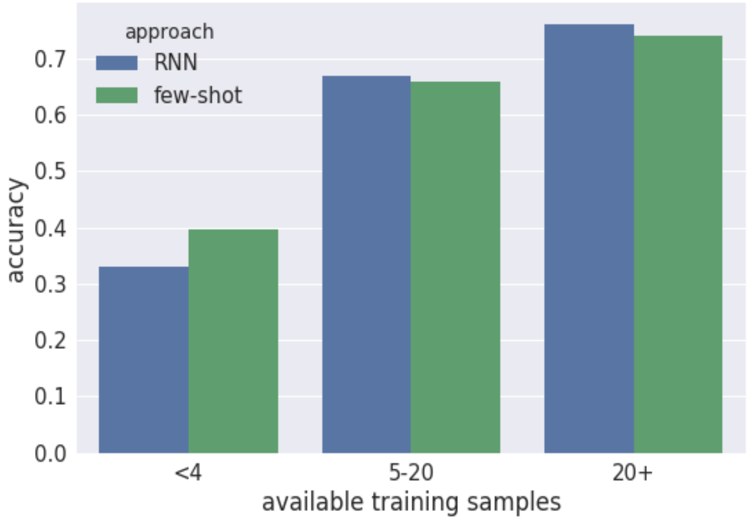

Figure 4.3: Accuracy per Available training samples. RNN vs FSL

Approach	Strict F score	Relaxed F score
FSL	**0.244**	**0.345**
RNN	0.239	0.327
Task Average	0.211	0.297

Table 4.4: SMM4H subTask 3 results

4.3.3 Evaluating FSL in SMM4H 2017 dataset

The evaluation of our approach in the previous section, showed promising results in the feasibility of the few-shot learning methods in the medical concept normalization domain. However, we wanted to evaluate our approach on a much larger dataset, where the number of available training samples would be large enough for more medical concepts and we would also be able to further analyze our results on the test set as the labels are known. In this context , the deep neural network approaches should be able to take advantage of it and demonstrate a better performance, based on our hypothesis. The SMM4H 2017 dataset, being the largest Twitter dataset available for medical concept normalization, consists of 3629 unique ADR mentions mapped to 507 MEDDRA codes.

In our 5-fold cross validation method we made sure that there was no overlap between the training set and the corresponding development and test set (as mentioned in the Dataset Description section). We evaluated our few-shot learning model by averaging the results from each fold. In order to determine the performance of the state of the art RNN and CNN on each fold, we averaged their performance on five different runs. The reason is that the random initialization of weights, slightly changed the final accuracy of the model in every run. In Table 4.5 we are presenting the most important hyperparameters of the two Neural Network models which were tuned in the development set folds of this dataset. In Table 4.6 the results of our 5-fold cross validation demonstrate that there is no significant difference in the performance of our few-shot learning approach compared to the best performing RNN, contrary to the workshop results presented in the previous section. We can assume, that the larger size of the

Hyper Parameters	**RNN**	**CNN**
Learning Rate	0.01	0.01
Epochs	20	40
Batch size	50	50

Table 4.5: NN Hyperparameters

Approach	**SMM4H 2017 Test. Set Accuracy**
CNN	0.42
RNN	0.561
FSL	0.565

Table 4.6: FSL and SOT 5 fold cross validation accuracy

dataset gave a boost to the performance of the neural network, compared to the one we used for training our model in the SMM4H 2019 shared Task. When analyzing

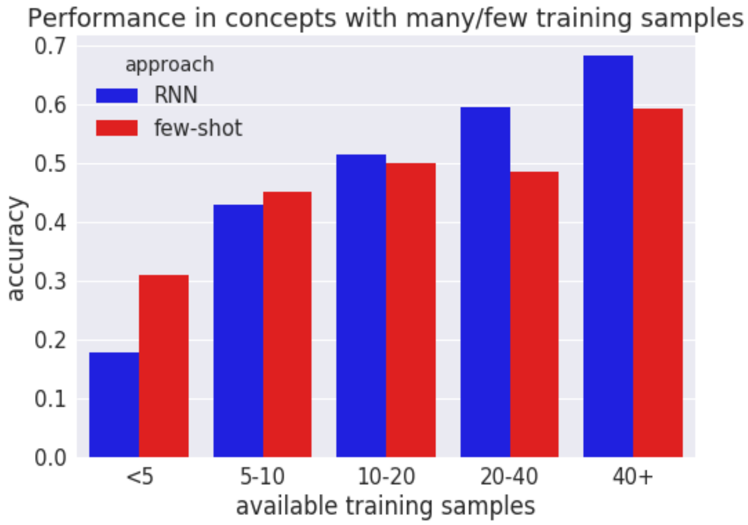

Figure 4.4: Accuracy per Available training samples. RNN vs FSL

the accuracy of the two best performing models as a function of the available training samples however, we can draw two basic conclusions similar to those of our initial experiment. First of all, our few-shot learning approach clearly outperforms the Recurrent Neural Network when the number of training samples is limited (less than 5), despite the fact that the overall performance of the two approaches is close. Secondly, as the number of the available training samples per concept increases, the RNN is constantly improving its accuracy, leading to a significantly better performance when we have more than 20 training samples per medical concept. The larger availability of training samples is also boosting the performance of the FSL but it is clear that the RNN can extract more complex and useful features in that case, than the simple cosine distance between the ADR vectors used by the FSL.

4.4 Hypothesis 2 Testing: Evaluating FSL in ensemble set-up

The results of our previous experiments proved that there is no one-size fits all solution to our problem. We can use a simple few-shot learning approach when we want a better performance in more rare medical concepts, but on the other hand if we want to be able to normalize the most usual Adverse Events that are present in user-generated text then probably using a deep learning model would demonstrate remarkable performance. Therefore, testing our initial hypothesis, that the proposed FSL model can effectively be used in an ensemble set up with the SOTA neural networks is of significant importance. In other words, this means that we have to examine if there is an effective way of discriminating at test time an ADR mention that belongs to a common concepts with many training samples,from a rare adverse event . In the following subsections we will describe our selected ensemble model and evaluate its performance compared to its two main components.

4.4.1 Selecting the optimal ensemble set-up

The basic requirement for our ensemble approach is to be able to determine with a high confidence whether an unlabeled sample should be classified with the RNN approach or with the Few-shot Learning approach. A quite straightforward way to do that is to generate a prediction from both models and then trust the most confident one. The level of confidence for the FSL approach can be determined by the cosine similarity of the sample's nearest neighbor and the confidence of the neural network model can be determined by the returned probability of the predicted label.

Another variation of this approach would be to train the Neural Network, only for predicting those classes for which the training data is sufficient. Then, assuming again that the returned probability reflects the level of confidence of the model, we can reject predictions lower than a probability threshold and consider them unknown. In that case, we would use the FSL approach. This alternative seems to be more promising for one reason. We would reduce the number of predicted classes among which the NN has to classify the unknown ADR mentions, that could possibly lead to further improvement of the classification accuracy in the well-known/ common concepts.

However, there seems to be a serious drawback in those two choices. Considering the softmax output probability of a neural network as a measure of confidence does not seem to be a valid hypothesis. As relevant research indicates [44, 10] deep neural networks can potentially produce high confidence outputs even when they are classifying totally unrecognizable input samples. To test that we performed the following experiment. We trained our state of the art RNN only with those medical concepts that have more than 5 training samples available. After training our model for 20 epochs, we made a prediction for each sample in the development set and visualized the output probability. Then we grouped the results based on the concept they belonged to (if it was rare or common) and based on the prediction validity (correctly classified or erroneously classified). As we can see from the results in the table above, the returned probability of the model is high even when it is predicting a previously unseen class. The average softmax output probability for those cases is 0.79 indicating a relatively high confidence of the model for those cases. In general, the correctly predicted concepts have higher softmax probability outputs but the gap is marginal. So setting a

No. of training samples	overall mean	wrongly predicted	correctly predicted
<5	0.79	**0.79**	0
5 to 9	0.73	0.56	0.84
10 to 19	0.84	0.732	0.91
20to39	0.67	0.46	**0.79**
40+	0.887	0.7	**0.95**

Table 4.7: RNN returned probability on SMM4H2017 development sets

threshold in the softmax output for recognizing commonly seen concepts from the rest is indeed a risk as relevant research indicates.

As an alternative, we can consider the prediction output of the few-shot learning approach as a confidence indicator. Our few-shot learning approach, is in fact a variation of the 1-NN classifier, where cosine distance is used as a distance measurement. Nearest Neighbor classifier, because of its nature, suffers from skewed class distributions. This means that, if a concept is very common in the training data, it will tend to dominate the labeling of the new samples increasing the number of False Positives in those common classes. In our case however, this nearest neighbor based approach still performs much better than a neural network does in rare concepts. In that sense, we expect that in rare concepts (the ones we are interested in), the FSL approach will have a relatively average recall but will achieve a high level of precision. In other words, our assumption is that if we trust our FSL approach only when predicting a concept with limited training examples (and in any other case use the RNN model) we will manage to achieve a better overall performance, as it is less likely that the FSL produces too many false positives in those Rare classes. In addition we will potentially increase the accuracy of the Recurrent neural network too, as we will significantly decrease the number of predicted classes.

4.4.2 Evaluating the FSL-RNN ensemble

An overview of the ensemble approach, as described in the previous section can be seen in the Figure 4.5 below. In our ensemble we use the FSL to find the most similar known vector representation for the unlabeled input ADR mention and if the predicted medical concept belongs to one of the Rare categories we assign this label directly. In any other case, we trust the RNN prediction. The RNN, is only trained on medical concepts that belong to the 'common' category, as they have sufficient training samples. In that way, we aim to reduce the number of predicted classes and see if this can give a boost to the performance on the rest.

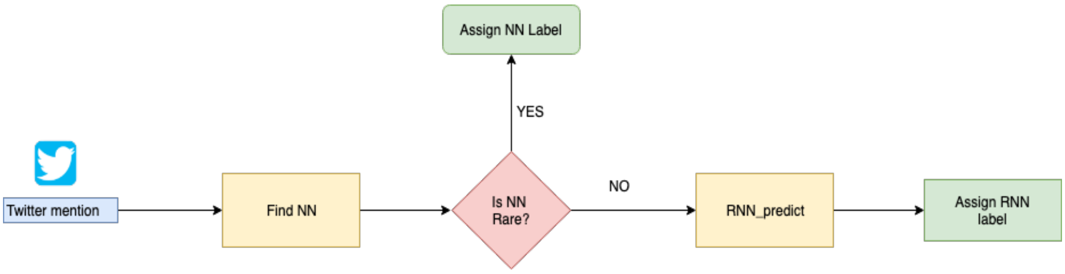

Figure 4.5: Ensemble of RNN and FSL

The threshold between the rare and the common concepts was determined based on the performance of the two approaches on the development set of each fold in order to avoid overfitting on the test set. In all 5 cases, we concluded considering all medical concepts with less than 5 available training samples as 'Rare' and all the rest were considered as common. After that discrimination, the remaining common classes that we trained the RNN with, were reduced to 115 from 507, which is a remarkable

Ensemble Hyperparameters	**Value**
RNN Learning Rate	*0.01*
RNN Batch size	*50*
RNN Epochs	*13*
Training samples Threshold for Rare concepts	*5*

Table 4.8: Ensemble approach Performance against SOT and FSL

Approach	**SMM4H 2017 Test. Set Accuracy**
RNN	*0.561*
FSL	*0.565*
Ensemble	***0.591***

Table 4.9: Ensemble approach Performance against SOT and FSL

change. From Table 4.9 we can see that the performance of the ensemble approach is significantly higher than both the FSL approach and the SOT RNN. This approach proves to be quite effective in discriminating concepts that should be classified with the FSL approach from the ones that are quite common and are suitable for a deep learning classification model. This is clearly demonstrated in Figure 4.6, where the performance of the ensemble approach among the different concepts in plotted. As we can observe, the ensemble approach keeps the same level of effectiveness in the rare medical concepts (the ones that have less than 5 training samples) and then manages to achieve a performance comparable to the RNN accuracy on common concepts with a large number of training examples. The above confirms our hypothesis that the number of false positives of the FSL is limited in rare medical concepts as the number of representatives of those cases in the vector space is limited. Apart from that however, we analyzed the performance of the RNN in the ensemble approach to examine

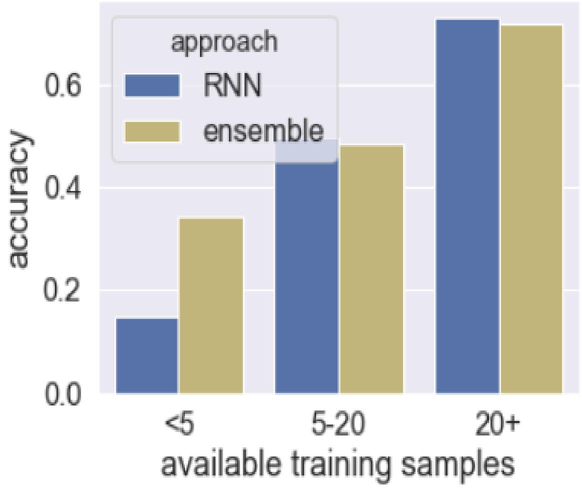

Figure 4.6: Accuracy per Available training samples. RNN vs Ensemble

whether the reduction of classes among which a prediction is made actually improved its performance in the remaining classes. The following Figure (4.7) is indicative. As we can see, there is a small improvement in the performance of the RNN when we reduce the number of classes, however this only happens in the middle bin where the available samples are between 5 and 20. In the last bin, where we include all concepts with 20 or more training samples the performance remains the same. This indicates two things. First of all that the reduction of the available classes helps the neural network improve its performance in cases of uncertainty, while its does not really affect the performance of the network in cases where the large availability of training data allows the network to predict those classes with high confidence.

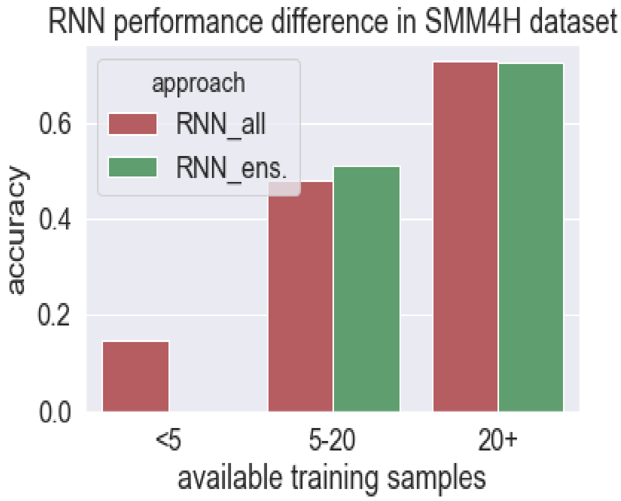

Figure 4.7: RNN vs RNN-Ensemble: Accuracy per available training sample

4.5 Hypothesis 3 testing: Adding prior knowledge

In the previous sections we made all those experiments that would allow us to draw useful conclusions about the feasibility of a few-shot learning approach in the field of medical concept normalization. From our perspective, we proposed this approach to fill in the lack of effectiveness that deep neural networks have in normalizing rare Adverse Drug Reaction concepts.

In this subsection we will go a step further. As mentioned in the previous chapter of this report, our hypothesis is that a few-shot learning approach like the one proposed, would be useful for normalizing medical concepts that are not present at all in the training data. To achieve that, we will use-prior knowledge from the KB in order to create prototypes of the medical concepts for which we have no training data. Medical Knowledge bases, like MEDDRA or UMLS associate each medical code with one or more medical terms representing them. Despite the fact that those terms belong to a standard medical vocabulary which is different than user-generated text, it can be useful in cases when we have very few or limited number or representatives of this concept from user-generated TWITTER data.

However, the way to use prior knowledge from the Knowledge-Base is not straightforward. The largest medical knowledge base, UMLS, associates each concept unique identifier (medical concept) with more than one synonyms of this term. For instance the concept unique identifier '*C0020517*' is associated with the term '*Hypersensitivity*' but also with the synonymous terms '*allergic reaction*', '*allergy disorders*' etc. MEDDRA KB which is the one used for the annotation of our current dataset, is a subset of UMLS. Therefore, we there is a direct mapping between the MEDRRA codes in our dataset and all those synonym terms associated with that code. Those synonyms are belonging to different categories based on their identity. For instance ACR terms include acronyms of a medical concept while CHV (consumer health vocabulary) include concepts that are mainly used by patients with non medical expertise to describe a medical concept. In order to determine, which of those families is more effective to use it as a concept representative for our FSL approach, we performed the following experiment.

4.5.1 Selecting the correct vocabulary

We considered no training data available for this task in order to evaluate to what extent prior knowledge can be used to normalize unseen ADR twitter mentions to MEDDRA codes in our SMM4H 2017 development set folds. We considered three different vocabularies of prior knowledge to find the optimal one for our task. Initially we only considered the Preferred Term (PT) for every MEDDRA code, which is a common formal medical term associated with a concept. As a second choice, we considered all the corresponding CHV synonyms from the UMLS Knowledge Base, as consumer Health Vocabulary has a higher chance of being more similar to user-generated text that we are trying to normalize. Finally, we also tried to consider the whole corpus that is available for a medical concept in UMLS without making any distinctions on the category of synonyms it belongs to. The results of this experiment as well as a comparison of the FSL approach when using the training data instead is presented in Table 4.10. From the presented results, it is interesting to mention that CHV (consumer Health vocabulary) achieves the best which can be explained by the fact that it is closer in nature to user-generated text mentions. On the other hand, even if we enrich CHV synonyms with all other related terms for a medical concept it does not seem to further improve

Vocabulary	FSL Accuracy
PT+CHV	***0.3***
All synonyms	*0.3*
PT only	*0.22*
Twitter training data	***0.48***

Table 4.10: FSL Accuracy on SMM4H2017 Dev. Set with prior knowledge only. PT= Using Preferred Terms only . PT+ CHV= Using PT and Consumer Health Vocabulary. All synonyms=Using all terms associated with a UMLS code

the performance. Probably the cases where the larger number of synonyms is helpful is not more than the cases where those terms actually add noise to the FSL approach.

4.5.2 Adding prior knowledge to FSL approach

Taking into account the results of the above experiment, we mapped all the considered medical concepts in our dataset (507 MEDDRA codes) to their corresponding Preferred Terms (PT) and their CHV synonyms from UMLS. We used this prior knowledge as additional training data apart from the TWITTER data that we were already using to see how the FSL and ensemble approach performance is affected. The results of our 5-fold cross validation evaluation are demonstrated in Table 4.11 and 4.12.

Approach	**Accuracy**
FSL	0.561
FSL+prior knowledge	0.5866

Table 4.11: FSL Performance difference with prior knowledge data

Approach	**SMM4H 2017 Test. Set Accuracy**
RNN	*0.561*
FSL	*0.565*
Ensemble	*0.591*
Ensemble + prior knowledge	***0.615***

Table 4.12: Ensemble approach Performance with prior knowledge data

4.5.3 Robustness of FSL to random noise

In our previous experiment, the reported performance of the ensemble approach can be considered overoptimistic. In a real case scenario, we do not know in advance which medical concepts will show up at test time. This means that if we add prior knowledge from concepts that are not present in the training set, we run the risk of adding random noise to our FSL model as it is obvious that some of those medical concepts will never be present at test time. In that sense, it is worth measuring to what extent our ensemble model is affected by adding random noise from medical concepts that are not present at test time. For this reason, apart from adding prior knowledge from the 507 MEDDRA codes that exist in the SMM4H 2017 dataset, we considered another 500 MEDDRA codes from the SIDER4 [2] collection of Adverse Events. Having almost doubled the considered medical concepts, we can see that the performance of the ensemble model is reduced, however it remains higher that the accuracy of the model when we did not use any sort of training data.

[2] http://sideeffects.embl.de/

Approach	SMM4H 2017 Test Set Accuracy
RNN	*0.561*
FSL	*0.565*
Ensemble	*0.591*
Ensemble(+prior knowledge)	*0.615*
Ensemble(+extended prior knowledge)*	*0.607*

Table 4.13: Ensemble approach Performance with prior knowledge data and random noise

4.6 Qualitative analysis

In this section, we will perform a qualitative analysis of our Few Shot Learning approach performance on the SMM4H 2017 dataset. The purpose of the qualitative analysis is to get insights about 'where' and 'why' this approach fails or succeeds and highlight all those underlying properties which are hard to digitize without losing any meaning. The most significant findings of our qualitative analysis can be summarized in the following experiments.

- Initially we tried to visualize ADR mentions that were correctly normalized to their corresponding medical concepts in order to identify useful patterns about the cases that our approach succeeds. As we mentioned in the introduction of this report, traditional string matching techniques do not perform well in the context of social media, as the use of language is totally different.

Unlabeled ADR	Predicted Concept
'Pulled an all nighter'	10022437 Insomnia
'sweat like a thunder cloud'	10020642 Hyperhidrosis
'walking in fog'	10041349 Somnolence
'grind my teeth sooooooooo'	10006514 Bruxism

Table 4.14: ADR mentions with no string overlap with their labels (concepts)

As demonstrated in the examples of Table 4.14, at first glance it seems that our approach is capable of normalizing medical concepts that have zero string overlap with their corresponding medical concepts.

- However when we tried to explore the 'closest' ADR vector representation (in terms of cosine distance) from the training,which actually determines the label of an unseen text mention, we came across various cases where the unseen ADR text mentions of the test data, are almost identical or very similar to ADR mentions in the training set. Please note here, that we have removed all duplicates between the training set and the test set of each fold, in order to reduce bias in our performance estimation during the experiments. As you can see in Table

4.15, there are several ADR phrases whose only difference is the word order or one token which can be considered a stop-word or a non significant word in terms of the semantic information it captures. It is obvious that in cases like that, a simple term matching technique that would use the training data, could possibly achieve a similar performance. However, it is also important to say that users in social media tend to express themselves in very different and informal ways, but they also tend to use similar expressions with each other.

Unlabeled ADR	Predicted Concept	Closest training data ADR
'13 hours of sleep'	10020765 Hypersomnia	*'sleep for 15+ hours'*
'never gonna go to sleep'	10022437 Insomnia	*'never gonna sleep'*
'grind my teeth sooooooooo'	10006514 Bruxism	*'grind my teeth'*
'food doesn't look appetizing'	10061428 Decreased Appetite	*'don't eat'*
'shuddering'	10044565 Tremor	*'jolting'*
'lack of nutrition'	10061428 Decreased Appetite	*'lack of hunger'*

Table 4.15: Examples of correctly normalized ADRs and their nearest neighbors from the training set

Apart from identifying patterns which highlight the strengths of our FSL approach we also performed some error analysis. We can clearly identify three different cases where our approach seems to produce the majority of the incorrectly normalized ADRs.

- First of all, we can see many examples of ADRs that were not normalized to their assigned medical concept, based on the annotator's choice, however our model selected semantically similar concepts in the classification procedure. The phrase *'kills my sex drive'* for instance in Table 4.16, is normalized to the medical concept *'Loss of Libido'* while the ground truth is *'Libido decreased'*. Despite the fact that they have almost the same meaning, those medical concepts represent two different entities in MEDDRA KB. This indicates a possible improvement direction in the future design of a medical concept normalization system. Grouping semantically similar medical concepts in a KB would enable a social media phrase to be normalized to more than one medical concepts, as they represent the same Adverse Event. In other words, we should consider the medical concept normalization problem as a multi-label classification task.

ADR mention	Predicted Concept	Ground Truth
'had me hooked'	Drug dependence (10013663)	Withdrawal syndrome (10048010)
suppressingg the fuck out my hunger'	Hunger (10020466)	Decreased appetite (10061428)
'kills my sex drive'	Loss of libido (10024870)	Libido decreased (10024419)
'weird ass dreams'	Nightmare (10029412)	Abnormal dreams (10000125)

Table 4.16: Examples of incorrectly normalized ADRs

- Of course, not all our erroneously normalized medical concepts belong to this category. It seems that our FSL model, by aggregating the individual word embeddings of an ADR phrase, fails in some cases to take into account significant properties of textual data, like word order or negation. As you can see in Table 4.17 the text mention *'never going to lose weight'* is erroneously normalized to *Weight Increased* because its closest neighbor in the vector space is 'lose so much weight'.

Unlabeled ADR	Predicted Concept	Closest ADR from training data	Ground Truth Concept
'1-2 hours of sleep'	Hypersomnia (10020765)	*'13 hours of sleep'*	Insomnia (10022437)
'never going to lose weight'	Weight decreased (10047895)	*'lose so much weight'*	Weight increased (10047899)
'high blood pressure'	Blood pressure decreased (10005734)	*'blood pressure low'*	Hypertension (10020772)
'never ate'	Increased Appetite (10021654)	*'ate'*	Decreased Appetite (10061428)

Table 4.17: Examples of incorrectly normalized ADRs

- Finally, some examples also indicated that adding a lexical normalization module like presented in [8] for our considered ADR text mentions would help our model avoid many Out of Vocabulary exceptions (OOV). When a token is misspelled, our word embedding model does not recognize it, so it can not generate any vector representation for it. This will make the normalization step impossible even if the ADR mention and the corresponding medical concept are almost identical strings like the examples in Table 4.18

Unlabeled ADR	**Ground Truth Concept**
'noappetite'	10061428 Decreased appetite
'withdrawels'	10048010 Withdrawal syndrome
'headachey'	10019211 Headache
'diarrhoea'	10012735 Diarrhoea

Table 4.18: Examples of incorrectly normalized ADRs due to OOV exceptions

4.7 Limitations

The last experiment presented in this chapter aims at pointing out the most important limitations of our approach. Based on theory, we expect our approach to have 3 basic disadvantages compared to the state of the art approaches. Those have to do with its performance when an ADR mention is long, so the aggregation of individual word embeddings is not very effective, its efficiency in terms of execution time and finally its inability to take into account the sequential nature of the text sequences. For the last one, we already provided the reader with an insight in the qualitative result analysis section. The other two are analyzed more extensively here.

4.7.1 Efficiency of the FSL approach

Apart from the evaluation of the few-shot learning approach in the effectiveness level, it is important to compare its efficiency with the state-of-the-art neural network. This comparison is not straightforward however. A neural network, requires a time consuming training process, which varies based on the size of the training data as well as the number of trainable parameters of the neural network. On the other hand, a few shot learning approach like the one that we present here, does not require any training. The only step that has to be taken is to embed the training data in the embedding space.

However, at test time, a neural network can be extremely efficient as the only action required to normalize a new sample is to perform a set of mathematical operations on the input and classify it based on the output of this operation. On the contrary, the few-shot learning approach is trying to find the nearest neighbor representation of the unknown samples, so it has to compute the cosine distance between every unlabeled sample and each one of the labeled training samples. In the Table 4.19 above we are demonstrating the average execution time of both methods for each one of the

Approach	**Execution Time** (sec)
FSL	*90*
RNN	*0.5*

Table 4.19: Execution time of SOT RNN and FSL in 400 sample test set

SMM4H 2017 test set folds. As we can see, the average execution time of the few-shot learning approach for normalizing 400 ADR mentions in an intel core i-7 CPU with 8 GB RAM is 90 seconds. The huge difference of the efficiency between this approach and an FSL approach indicates that in certain use cases, like real-time applications , a FSL approach like that would be hard to use. It is worth mentioning however, that our implementation for finding the nearest neighbor in terms of cosine distance,was not performed using a built in PYTHON library, as we needed to implement this function ourselves. Therefore, improving our implementation or adding parallel processing of those calculations could potentially reduce the execution time in more acceptable levels.

4.7.2 Effectiveness of FSL approach in long ADR mentions

As we mentioned in the previous chapter of this report , we selected an approach where we would simply aggregate the different word embeddings of an ADR to create a fixed-size vector representation, because of the fact that an Adverse Event is usually described within a few tokens rather than a whole sentence or paragraph. Our hypothesis was that aggregation of individual word embeddings can still capture the semantics of a short phrase, but the longer the phrase becomes the more noisy this representation will be. For this reason we measured the performance of our approach as a function of the ADR mention length to identify to what extend the mention length affects the model's performance. Surprisingly, as Figure 4.8 indicates the Accuracy of the few-shot learning approach does not seem to be affected by the mention length in ADRs between 1 to 6 tokens. The decrease in performance is obvious only for mentions that are described with more than 6 tokens, but the this is a very small percentage of the whole data (less than 10 %). On the other hand, if we add the RNN performance as a baseline, we can clearly see that the difference in the achieved accuracy between the two models increases in favor of the RNN as the ADR mention length gets larger. The neural network is also affected by the length of the mention, however it is more robust as Figure 4.9 indicates. As an overall conclusion however, we could say that even in comparison to the RNN the performance of our simple FSL model is still competitive even when the mention length increases.

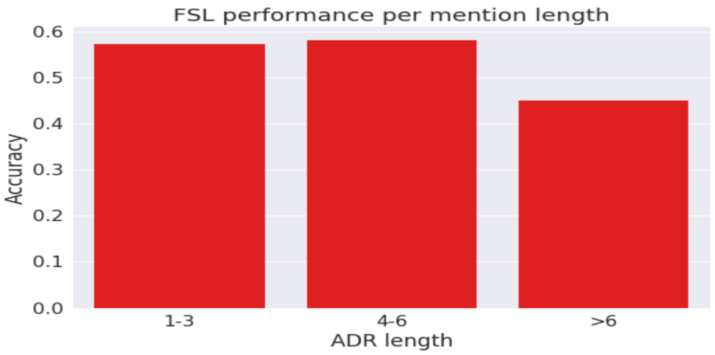

Figure 4.8: FSL Accuracy as a function of the ADR length

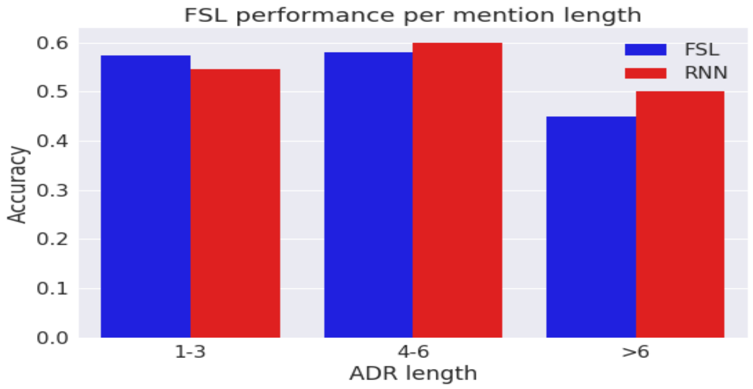

Figure 4.9: FSL vs RNN Accuracy as a function of the ADR length

4.8 Experimental Result Discussion

Having presented the results of various experiments in the previous subsection, we will now discuss the results of this evaluation process and conclude whether those results confirm or reject each one of our hypothesis about the proposed technique.

H1: Our Few-Shot Learning Approach will perform better than the state of the art in normalizing medical concepts with limited training samples. The state of the art Neural Networks will outperform the few-shot learning approach on classes (concept) with a large availability of training samples.

Our first hypothesis regarding our proposed few-shot learning approach is confirmed by the experimental results in our real world dataset. This is clearly indicated by the result analysis of the SMM4H 2017 data in Figure 4.4 as well as the same analysis done in the development set of the SMM4H 2019 shared Task (Figure 4.3). It is proven that a simple non-parametric approach, based on a simple aggregation of the individual word embeddings of an ADR phrase, not only projects the ADR mentions in an embedding space where (semantically) similar from dissimilar ADRs can be distinguished, but is also capable of performing much better than a complex NN model in cases with limited training examples as the NN has a clear tendency to overfit in those cases.

On the other hand as the number of the available training examples per class increase the deep neural network model is superior, however the FSL approach performed surprisingly well even in those cases. It is indicative that in the SMM4H 2019 Development set, the difference between the state of the art RNN and the FSL is marginal in medical concepts that have at least 20 unique training samples. This means that our proposed approach is also taking advantage of the availability of training data. However, in larger datasets like the SMM4H 2017, which is approximately 3 times larger than the SMM4H 2019 dataset and also there is zero overlap between the training set and the test set data, it is clearly demonstrated (Figure 4.4) that the NN is capable of extracting more useful features taking advantage of the variation in the training data in commonly seen medical concepts.

As an overall best solution however we can not conclude that one of those approaches is proffered over the other. It can be seen that the kind of the dataset plays an important role in the approach that is more suitable in each case. A smaller dataset , where the majority of the data will belong to medical concepts (labels) with limited training samples, is a good fit to a few-shot learning approach. If however, we have to do with a more extensive data source where the majority of the considered medical concepts has sufficient training samples to train a neural network then, a deep neural network could demonstrate a remarkable performance. Apart from the training data available, the use case of the problem also plays an important role in the model selection. As mentioned earlier, the execution time needed by our proposed method is significantly higher than the time needed by a trained RNN to classify unseen samples. Therefore, in cases of real time applications for instance, a DNN model would be a better choice than a NN based approach.

H2: We can combine the proposed FSL approach with the current SOTA deep neural networks to achieve a more robust performance among the different imbalanced classes.

Based on the experimental evaluation of the corresponding section, we can conclude that the proposed FSL approach, can not only be used to solve the limited training data problem is certain classes, but it is also capable of being used in a complementary set up with a deep neural network. In that case, the ensemble model is capable of clearly outperforming the deep neural network alone as well as the FSL baseline, and demonstrates a more robust performance among the difference medical concepts. A system like that can avoid overfitting by trying to generalize to classes it has barely seen before, while at the same time it can demonstrate a remarkable performance in classes where the variation within the class training examples is large. On the other hand, when the amount of annotated data is so small that most of the considered medical concepts have few representatives, the FSL technique can also demonstrate perform as a standalone system. This is clearly indicated in our 4.3.2 Evaluation section.

H3: Prior knowledge can be used as an alternative data to normalize medical concepts when no training data is available.

The above hypothesis, taking into account the results in Table 4.10, can not be confirmed 100 %. As we can see, prior knowledge has limited ability of being used as a representative of a medical concept in the user-generated context. It is obvious that the difference in the use of language between medical KB and social media is affecting the performance of the FSL proposed method. On the other hand, we can not ignore that it is a very useful source of additional data, as the combination of social media training data and prior knowledge data, increases the performance of the FSL approach as a standalone system, as well as the performance of the RNN-FSL ensemble which we presented previously.

Chapter 5

Conclusions and Future Work

In this final chapter we include the main conclusions of our work,discuss our decisions and critically reflect on the whole work we did. Finally we pose our suggestions for future researchers. The structure of this section is as follows. In section 5.1 we will summarize our findings on the research questions that we posed in the Introduction of our research, while in section 5.2 we are discussing our main decisions taken along the way and how those affected our work and our findings. Future work suggestions are included in Section 5.3.

5.1 Conclusions

The main research question that we pose in this work is '*How can we link medical entities mentioned in user-generated text, to their corresponding entities in an existing Knowledge Base?*'. The answer to this high level research question can be given by the findings of the following three research sub questions that we introduced in the earlier phases of our work.

RSQ1: *What are the state of the art methods, for linking medical text entities to entities KB entities?*

In order to answer this research question we performed a systematic literature study and collected the most significant scientific work in the domain of medical entity linking/ medical concept normalization. The findings of this survey indicate that the state of the art techniques for normalizing medical entities to a standard Knowledge Base vocabulary, are based on the semantic representation of the text entities into vectors and the use of deep neural networks on top of them so they can extract useful features and normalize them to the correct output class (medical concept). Deep learning techniques, especially in user-generated text, are reported to outperform traditional rule-based or string matching techniques , as well as all the other supervised and unsupervised Machine Learning based approaches that are present in literature.

However, after analyzing the limitations of those state of the art approaches, we realized that the nature of the medical concept normalization problem poses the limitations itself. This is because as expected some medical concepts like Adverse Events, Diseases or Symptoms are more common than others. For those medical concepts we usually have (or at least can easily acquire) sufficient annotated data to train a neural network model. On the other hand, the majority of medical concepts are not very commonly mentioned in user-generated text for instance. Therefore, because of this class imbalance issue, most deep neural network researchers try to normalize a large number of medical concepts, among which only a small percentage has sufficient training samples. This leads to a remarkable performance on the common concepts but the accuracy in normalizing more rare classes is disappointing. As a consequence, we have to mention here that accuracy which is the only metric used in relevant research to evaluate medical concept normalization techniques is not enough to give us the necessary insights about the performance of a system. In our work we mostly analyze the accuracy of our systems in different families of classes depending on the availability of training data.

RSQ2: *How can we address the drawbacks and limitations of the current state of the art techniques in normalizing ADRs in user-generated text?*

In our work, we focused on building a technique that could fill in the gaps of the state of the art deep learning approaches and manage to improve the classification accuracy in concepts where the annotated data is limited. In machine learning theory,an algorithm that is able to perform a classification task with only a few 'shots' from each class is called '*few-shot learning*' algorithm.Our Few- shot learning approach tries to simulate the human learning process, where a human is able to identify an object or a living creature that he/she has only seen once again in the past, by recognizing the similarity of a new unseen object with what he or she has been shown in the past. More specifically, our few-shot learning algorithm tries to create an embedding (representation) of an Adverse Drug Reaction mentioned in user-generated text,in an embedding space where semantically similar from semantically dissimilar representations would be easily identified with the use of a similarity metric. To create this vector representation of our medical concepts (ADRs), we were based on pre-trained word embedding models, which are also successfully used to create the representations of the input ADRs in the state of the art deep learning models. Since, the largest percentage of the Adverse Drug Reaction entities in user-generated text are composed of multiple tokens, we had to find a way to represent the whole ADR phrase as a fixed-size vector, using the individual token vectors of that phrase. Taking into account that we wanted to achieve the highest possible performance in concepts with a limited number of training samples, we decided to select a very simple and straightforward way of creating this fixed-size vector representation in order to avoid overfitting. After some experimentation, we concluded that element-wise addition of the individual word embeddings of an ADR phrase could still capture the semantics of the phrase, in such a way that the cosine similarity (cosine distance between the phrase vectors) between two semantically similar ADRs was high.

RSQ3: *How effective is our proposed approach in linking ADR mentions from user-generated text compared to the state of the art approaches?*

To answer this research question, we performed extensive experimental evaluation of our approach in real world data from TWITTER. To test the effectiveness of the FSL approach we created an embedding for each ADR in the training data, and then we assigned a label to all the unknown test samples based on their nearest labeled neighbor from the training set. Our findings indicate that this simple approach is capable of demonstrating a superior performance in concepts with limited training examples, compared to the state of the art Deep learning techniques, which were reproduced on the same data. Apart from outperforming the state of the art in rare medical concepts, the FSL approach demonstrates a very competitive performance in the task in general. The overall performance of a reproduced RNN with a single GRU layer and our FSL approach is almost identical in the largest available TWITTER dataset for medical concept normalization, while the FSL is outperforming the RNN when evaluated on a smaller dataset. The reason , not surprisingly , is that a deep neural network has in general a much better performance in all medical concepts with a large variety between the training samples where it clearly outperforms the few-shot approach. This is an important finding, as it indicates that there is no one-size fits all solution to this problem. For this reason, we tried to evaluate our FSL approach as an alternative to the deep neural network model, when a rare medical concept has to be classified. We wanted to evaluate whether using our approach in combination with the SOT, to distinguish the rare from the common medical concepts is a feasible task. Indeed, due to the few false positives that the FSL approach produces in medical concepts with few representatives, we managed to build an ensemble model that demonstrates a quite robust performance among all different training concepts no matter how common or rare they are in the training data. As a final take away from our evaluation section, we could say that our FSL approach proves to be a feasible solution as a stand alone model in cases where the training data is small, as well as an alternative to the state of the art deep learning techniques in cases where the large availability of data in some medical concepts allows them to demonstrate remarkable performance on them.

5.2 Discussion

In this subsection we will discuss and reflect our decisions made through the whole research process.

First of all, based on the weakness of the current state of the art approaches we decided to evaluate a few-shot learning approach. The advantage of this decision was that we managed to achieve an improvement in the rare medical concepts, however it was quite likely from theory and proved to be true in practise that such an approach would not be able to compete a deep neural network in concepts where the training data was available in a larger scale.

Secondly, we decided to create a fixed size vector representation of our ADR text mentions by simply aggregating the individual token embeddings of the phrase using a pre-trained model. This led to a quite simple and straightforward to implement technique that does not suffer from a large number of trainable parameters as the SOTA approaches. However, we did not predict the fact that it would fail to capture the sequential nature of the data in several cases or to identify negation and hence misclassifying opposite medical concepts as being the same. This fact was mainly present when the length of the phrase was increasing. Considering different weights for tokens that express either negation (ie. 'never' , 'not') or range (i.e 'low' vs 'high'), could potentially reduce this risk in certain cases. Thirdly, we considered the medical concept normalization problem as a simple multi-class classification problem. As proved by our result qualitative analysis, there are several medical concepts in a knowledge base that are semantically similar (ie 'Arthalgia' and 'Joint pain'). Therefore, a more extensive exploration of the Knowledge Base and the identification of the relations between the different medical concepts could have given us better insights about the nature of the problem.

As a last remark we should mention the evaluation metrics used. In all related research on medical concept normalization in user-generated text, the only evaluation metric chosen was Accuracy. However it seems that because of the nature of the problem, even a normalization system that is able to effectively predict only a minority of very common medical concepts (i.e insomnia, headache) can demonstrate a competitive accuracy metric. Therefore, it would be worth using using precision, recall and F-score apart from the reported accuracy as a more reliable metric to evaluate the effectiveness of the approach across all the considered classes.

5.3 Future work

In this final subsection we will provide the reader with our proposals for possible future research and further extensions and improvement of our work. In earlier sections, we concluded about the feasibility, the strengths and the weaknesses of our proposed approach. Future research proposals will aim in the direction of improving the current limitations of our research method.

First of all, a possible direction for improvement would be to optimize the efficiency of the few-shot learning algorithm. As we demonstrated earlier, the proposed approach is extremely slow as it has to compute the cosine distance of each new test samples with all elements of the training data. Hashing techniques can be used to speed us this procedure, however a more research oriented direction for solving this would be to use the training data to create only one embedding (vector representation) for each medical concept. In that way, the number of prototypes that a new sample has to be compared with will be minimized.

Another possible research direction, that could potentially also reduce the lack of efficiency that we mentioned above, would be to evaluate different few-shot learning techniques in the medical concept normalization domain. For instance, the use of siamese neural networks in similar problems like job title normalization or question answering [34], where pairs of input text are classified as similar or dissimilar have demonstrated a remarkable result in their domains.

Finally, our model can be improved in the direction of separating the common and the rare medical concepts at test time in a more effective and efficient way. Using multiple binary classifiers, or neural networks with multiple sigmoid functions instead of the final softmax layer have been used in similar domains where multi-label classification or open-set classification [41] is considered.

Bibliography

[1] Alan R. Aronson and François Michel Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[2] Katherine Bailey and Sunny Chopra. Few-Shot Text Classification with Pre-Trained Word Embeddings and a Human in the Loop arXiv : 1804 . 02063v1 [cs . CL] 5 Apr 2018. pages 1–8, 2016.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[4] Anne Cocos, Alexander G Fiks, and Aaron J Masino. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821, 2017.

[5] C Combi, M Zorzi, G Pozzani, E Arzenton, and U Moretti. Normalizing Spontaneous Reports into MedDRA: some Experiments with MagiCoder. *IEEE Journal of Biomedical and Health Informatics*, 2018.

[6] Sébastien Cossin, Vianney Jouhet, Fleur Mougin, Gayo Diallo, and Frantz Thiessard. IAM at CLEF eHealth 2018: Concept annotation and coding in French death certificates. *CEUR Workshop Proceedings*, 2125, 2018.

[7] Arjun Magge Ashlynn Daughton Karen O'Connor Michael Paul Graciela Gonzalez-Hernandez. Davy Weissenbacher, Abeed Sarker. Overview of the fourth social media mining for health (smm4h) shared task at acl 2019. in proceedings of the 2019 acl workshop smm4h: The 4th social media mining for health applications workshop shared task. 2019.

[8] Anne Dirkson, Suzan Verberne, and Wessel Kraaij. Lexical normalization of user-generated medical text. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 11–20, Florence, Italy, August 2019. Association for Computational Linguistics.

[9] Ehsan Emadzadeh, Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez-Hernandez. Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology. *Proceedings of the American Medical Informatics Association Annual Symposium*, (June), 2017.

[10] Yarin Gal. Uncertainty in deep learning. 2016.

[11] Omid Ghiasvand and Rohit J Kate. UWM : Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. (SemEval):828–832, 2014.

[12] Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. Multimedia lab @ ACL WNUT NER shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153, Beijing, China, July 2015. Association for Computational Linguistics.

[13] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[14] Sifei Han, Tung Tran, Anthony Rios, and Ramakanth Kavuluru. Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter. *Proceedings of the 2nd Workshop on Social Media Mining for Health Research and Applications*, pages 49–53, 2017.

[15] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73 – 81, 2015.

[16] Dimitri Kartsaklis, Mohammad Taher Pilehvar, and Nigel Collier. Mapping Text to Knowledge Graph Entities using Multi-Sense LSTMs. pages 1959–1970, 2018.

[17] R.J. Kate. Normalizing clinical terms using learned edit distance patterns. *Journal of the American Medical Informatics Association*, 23(2):380–386, 2016.

[18] B. Kitchenham and S Charters. Guidelines for performing systematic literature reviews in software engineering, 2007.

[19] Barbara A. Kitchenham, David Budgen, and O. Pearl Brereton. Using mapping studies as the basis for further research â a participant-observer case study. *Information and Software Technology*, 53(6):638 – 651, 2011. Special Section: Best papers from the APSEC.

[20] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org, 2015.

[21] Robert Leaman, Rezarta Islamaj DoÇ§an, and Zhiyong Lu. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[22] Robert Leaman and Zhiyong Lu. Disease Named Entity Recognition and Normalization with DNorm. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, page 587, New York, NY, USA, 2014. ACM.

[23] Kathy Lee, Sadid A. Hasan, Oladimeji Farri, Alok Choudhary, and Ankit Agrawal. Medical Concept Normalization for Online User-Generated Texts. *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, pages 462–469, aug 2017.

[24] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics*, 18(Suppl 11), 2017.

[25] N Limsopatham and N Collier. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 1675–1680, 2015.

[26] N Limsopatham and N Collier. Normalising medical concepts in social media texts by learning semantic representation. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 2, pages 1014–1023, 2016.

[27] C. Lü, B. Chen, C. Lü, L. Qiu, and D. Ji. A multiple feature approach for disorder normalization in clinical notes. *Wuhan University Journal of Natural Sciences*, 21(6):482–490, 2016.

[28] C.J. Lu, D. Tormey, L. McCreedy, and A.C. Browne. *Enhanced lexsynonym acquisition for effective UMLS concept mapping*, volume 245. 2017.

[29] Adriaan M. J. Schakel and Benjamin J. Wilson. Measuring Word Significance using Distributed Representations of Words. 2015.

[30] Alejandro Metke-Jimenez and Sarvnaz Karimi. Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms. *ArXiv*, abs/1504.06936, 2015.

[31] Z Miftahutdinov and E Tutubalina. End-to-end deep framework for disease named entity recognition using social media data. In *IEEE 30th Jubilee Neumann Colloquium, NC 2017*, volume 2018-Janua, pages 47–52, 2018.

[32] Zulfat Miftahutdinov and Elena Tutubalina. KFU at CLEF eHealth 2017 Task 1: ICD-10 coding of English death certificates with recurrent neural networks. *CEUR Workshop Proceedings*, 1866(Icd), 2017.

[33] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[34] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. Learning text similarity with Siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany, August 2016. Association for Computational Linguistics.

[35] Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. Multi-task Character-Level Attentional Networks for Medical Concept Normalization. *Neural Processing Letters*, pages 1–18, 2018.

[36] Chongyu Pan, Jian Huang, Jianxing Gong, and Xingsheng Yuan. Few-Shot Transfer Learning for Text Classification With Lightweight Word Embedding Based Models. *IEEE Access*, 7:53296–53304, 2019.

[37] Jeffrey Pennington, Richard Socher, and Christoper Manning. Glove: Global vectors for word representation. volume 14, pages 1532–1543, 01 2014.

[38] Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207, 2015.

[39] Abeed Sarkerb, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M. Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. Data and systems for medication-related text classification and concept normalization from twitter: Insights from the social media mining for health (smm4h) 2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 10 2018.

[40] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline Needs More Love : On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. pages 440–450, 2018.

[41] Lei Shu, Hu Xu, and Bing Liu. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[42] Luca Soldaini. QuickUMLS : a fast , unsupervised approach for medical concept extraction. 2016.

[43] S.A. Stewart, M.E. Von Maltzahn, and S.S.R. Abidi. Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons. In *CEUR Workshop Proceedings*, volume 895, pages 63–77, 2012.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. 2013.

[45] Amir M Tahmasebi, Henghui Zhu, Gabriel Mankovich, Peter Prinsen, Prescott Klassen, Sam Pilato, Rob Van Ommering, Pritesh Patel, Martin L Gunn, and Paul Chang. Automatic Normalization of Anatomical Phrases in Radiology Reports Using Unsupervised Learning. 2018.

[46] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84(June):93–102, 2018.

[47] Yaqing Wang and Quanming Yao. Few-shot learning: A survey, 04 2019.

[48] Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77:29799–29810, 2018.

[49] Mo Yu, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Diverse Few-Shot Text Classification with Multiple Metrics. pages 1206–1215, 2018.

[50] Pierre Zweigenbaum and Thomas Lavergne. Hybrid methods for ICD-10 coding of death certificates. pages 96–105, 2016.

Appendix A

Paper

In this Appendix you can find our published paper in the fourth Social Media Mining for Health (SMM4H) workshop, part of ACL 2019 conference.

Give it a shot: Few-shot learning to normalize ADR mentions in Social Media posts

Manolis Manousogiannis
myTomorrows
Delft University of Technology
m.manousogiannis@mytomorrows.com

Sepideh Mesbah
Delft University of Technology
s.mesbah@tudelft.nl

Selene Baez Santamaria
myTomorrows
s.baez@mytomorrows.com

Alessandro Bozzon
Delft University of Technology
a.bozzon@tudelft.nl

Robert-Jan Sips
myTomorrows
r.sips@mytomorrows.com

Abstract

This paper describes the system that team myTomorrows-TU Delft developed for the 2019 Social Media Mining for Health Applications (SMM4H) Shared Task 3, for the end-to-end normalization of ADR tweet mentions to their corresponding MedDRA codes. For the first two steps, we reuse a state-of-the-art approach, focusing our contribution on the final entity-linking step. For that we propose a simple Few-Shot learning approach, based on pre-trained word embeddings and data from the UMLS, combined with the provided training data. Our system (relaxed F1: 0.337-0.345) outperforms the average (relaxed F1 0.2972) of the participants in this task, demonstrating the potential feasibility of few-shot learning in the context of medical text normalization.

1 Introduction

Team myTomorrows-TU Delft participated in subtask 3 of the 2019 Social Media Mining for Health Applications (SMM4H) (Davy Weissenbacher, 2019) workshop, which is an end-to-end task. The goal is, given a tweet, to 1) automatically classify tweets containing an adverse drug reaction mention; 2) extract the exact ADR mention; 3) normalize the extracted ADR to its corresponding Medical Dictionary for Regulatory Activities (MedDRA) code. The task is evaluated based on strict and relaxed F-score, precision and recall.

From an NLP perspective, this task poses a significant challenge as there is a large gap between the informal language used in social media and the formal medical language. Moreover, there is an absence of large annotated datasets, and datasets which are available often suffer from class imbalance. Illustrating this, Figure 1 provides an overview of the number of samples per class in the SMM4H task 3 dataset.

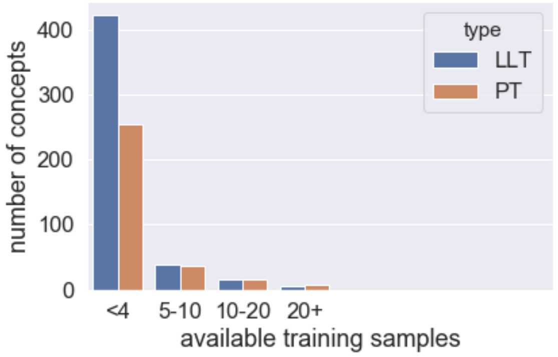

Figure 1: Available training samples per the medical concept present in the training data

Our end-to-end system consists of existing state-of-the-art for the first two steps. We focus our efforts on the third -normalization- step, which we formulate as a Few-Shot Learning problem (FSL), following the definition by Wang and Yao (Wang and Yao, 2019). In the following sections, we describe (1) the datasets that we worked on, (2) our approach in more detail and finally (3) our results and conclusions.

2 Data

2.1 Datasets

With the three subtasks, three manually annotated datasets were provided. All datasets contain tweets containing an ADR (positive) and without an ADR (negative). A brief overview of these datasets is provided in Table 1, but for more context we refer to (Davy Weissenbacher, 2019).

2.2 Preprocessing

The provided dataset for subtask 3 consists of ADR mentions, annotated with their corresponding MedDRA code. In the hierarchy[1] of MedDRA,

[1] https://www.meddra.org/how-to-use/basics/hierarchy

Task	**Training data**	
	#Positives	*#Negatives*
1	2374	23298
2	1212	1155
3	1212	1155

Table 1: Statistics of the training data used for task 1, 2 and 3

one Preferred Term (PT) is linked to one or more Lower Level Terms (LLTs) which are more specific descriptions of the related concept.

The provided dataset contains a mix of PTs and LLTs, mapping the 1212 ADR mentions to more than 500 different codes. Observing that the evaluation of the workshop task is performed on PT level, we map all annotations to the corresponding PT, as a preprocessing step. After this preprocessing step, the 1212 training mentions are mapped to 319 MEDDRA codes. Figure 1 provides an overview of the class distribution before and after preprocessing.

2.3 Prior Knowledge

In the training set for subtask 3, 149 out of the 319 MEDDRA codes that are present in the dataset (46.7%) have just one available training sample, while 254 (79.6%) have less than five training samples. To deal with the scarcity of samples, we create a prior knowledge dataset considering the 319 MEDDRA PTs in the training data. This dataset consists of the preferred names provided by the MEDDRA vocabulary and their corresponding preferred names in the Consumer Health Vocabulary (CHV), as mapped by the UMLS. The resulting dataset cointains 1,854 preferred names for the 319 MEDDRA codes.

3 Method

Our contributions focus on the normalization step, linking ADRs to their corresponding MEDDRA code. However, to be able to perform an end-to-end evaluation, we use existing state-of-the art techniques for subtask 1 (Sarker and Gonzalez, 2015) and 2 (Cocos et al., 2017), which we train on the workshop datasets [2].

The state-of-the-art approach for medical concept normalization in user-generated text is deep neural networks (Limsopatham and Collier, 2016) which outperform traditional methods, when sufficient training data are available.

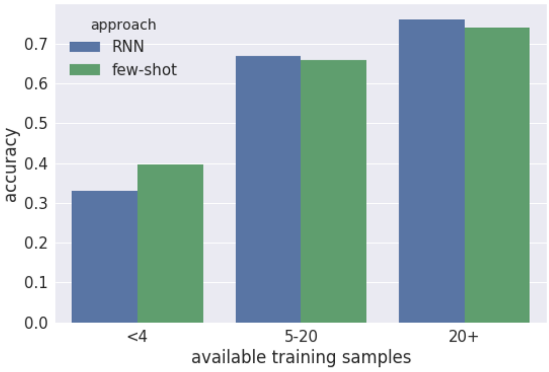

Figure 2: Accuracy per number of training samples.

We trained both the CNN and RNN described by (Limsopatham and Collier, 2016) on the dataset for task 3, finding that the RNN has the best performance. On closer observation (and not surprisingly), we found that the accuracy of the RNN drops when fewer samples are available in the training data, as depicted in figure 2.

To deal with this drop in performance, we propose an embedding-based classifier that compares the ADR extracted mention to its 1-Nearest Neighbour on a vector space containing a) representations of the ADR mentions in the training data and b) representations of the prior knowledge dataset. Our intuition is that the embedding-based binary classifier would perform better on classes with a low number of samples, whereas an RNN would perform well on classes with higher sample numbers.

To create our embedding-based classifier we employ the pretrained Google News Word2Vec model (Mikolov et al., 2013). Using this model, we create vector representations for the ADR mentions in our training data[3]. Similarly we create vector representations for the mentions gathered in our prior knowledge dataset. At test time, we employ the same Word2Vec model to create a vector representation of the unseen ADR mention. Using a 1-Nearest Neighbour (with cosine similarity as distance metric), we then select the corresponding MEDDRA concept. Figure 2 shows that this model indeed seems less sensitive to low sample numbers.

[2] For task 1, we trained using the suggested settings, assigning 3:1 class weight favouring the ADR class. For task 2, we trained using the pre-trained-fixed setting.

[3] for mentions of more than one token we added the vectors

Technique	Relaxed			Strict		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
RNN	*0.318*	*0.337*	*0.327*	*0.232*	*0.246*	*0.239*
FSL	**0.336**	**0.355**	**0.345**	**0.237**	**0.252**	**0.244**
RNN+FSL (1)	*0.328*	*0.347*	*0.337*	*0.23*	*0.244*	*0.237*
RNN+FSL (2)	*0.331*	*0.35*	*0.34*	*0.235*	*0.249*	*0.242*
Task 3 AVG	*0.29*	*0.311*	*0.297*	*0.205*	*0.224*	*0.211*

Table 2: Relaxed and strict Precision/Recall/F-score for RNN, FSL, RNN+FSL (1) and (2) and the average score of all the participated team in task 3 (Task 3 AVG)

For our experiments, we use 4 systems: (1) RNN: the RNN proposed by (Limsopatham and Collier, 2016), trained on the both prior knowledge and the training set (which provides the best performance), (2) FSL: our 1-NN based on a combination of prior knowledge and the training set, (3) RNN+FSL (1): an ensemble of the RNN trained on only the training set and the FSL based on training + prior knowledge, and (4) RNN+FSL (2): an ensemble of the RNN trained on the training set and prior knowledge and the FSL based on training + prior knowledge. For our ensembles, we trust the model with the highest confidence (we used the cosine similarity for the 1-NN model to represent confidence) in case of disagreement.

4 Results

Our results are summarized in Table 2. Despite the fact that the RNN+FSL performed better in our development set, it did not generalize in the test data. On the test and evaluation data, FSL outperformed all the other techniques and achieved a 0.345 relaxed F-score and a 0.244 strict F-score which are above the average performance achieved in this task by all participants (i.e. Task 3 AVG).

5 Conclusions

In this paper, we describe our approach in subtask 3 of the SMM4H shared task for normalization of Adverse drug reaction mentions in Twitter posts. Our few-shot learning approach performs above the average in this task and hence we believe it to be a promising approach in cases where the amount of training data is limited.

As future work, we will focus on the discrimination between the ADRs that belong to one of the 'commonly seen cases' (classes with sufficient training data) from the 'rare cases' (classes with insufficient training data). This will allow us to efficiently combine a deep neural network with a few-shot learning approach into a more robust system that successfully links ADR tweet mentions into its MEDDRA codes.

References

Anne Cocos, Alexander G Fiks, and Aaron J Masino. 2017. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Arjun Magge Ashlynn Daughton Karen O'Connor Michael Paul Graciela Gonzalez-Hernandez. Davy Weissenbacher, Abeed Sarker. 2019. Overview of the fourth social media mining for health (smm4h) shared task at acl 2019. in proceedings of the 2019 acl workshop smm4h: The 4th social media mining for health applications workshop shared task.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1014–1023.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Yaqing Wang and Quanming Yao. 2019. Few-shot learning: A survey. *arXiv preprint arXiv:1904.05046*.